Evolution in Sundaland: insights from comparative phylogeography of *Rattus* and *Sundamys* rats

MIGUEL CAMACHO SÁNCHEZ

ESTACIÓN BIOLÓGICA DE DOÑANA - CSIC

UNIVERSIDAD PABLO DE OLAVIDE

2017

Cover: *Sundamys infraluteus* in Gunung Alab, Sabah, Borneo. Credits: Arlo Hinckley

"pogun Sabah, tulun om
dupot, toput do tinaru deet olidang
kumaa'd kotolunan"

TESIS DOCTORAL

Evolution in Sundaland: insights from comparative phylogeography of *Rattus* and *Sundamys* rats

Memoria presentada por **Miguel Camacho Sánchez** para optar al Grado de Doctor por la Universidad Pablo de Olavide.

Fdo. Miguel Camacho Sánchez

Directora:

Dra. Jennifer A. Leonard

Estación Biológica de Doñana-CSIC

Tutora:

Dra. Martina Carrete

Estación Biológica de Doñana-CSIC

Universidad Pablo de Olavide

Summary

The tropical bioregion of Sundaland, Southeast Asia, is a major hotspot of world biodiversity, including mammal biodiversity. Its complex geological history lays out an excellent scenario to study evolution. However, the interpretations of Sunda biogeography patterns for this group has been limited due to (1) uncertainties in taxonomy, and (2) biased and incomplete sampling of certain regions (e.g. Sumatra). I combined a taxonomic approach followed with extensive sampling in the field and from natural history collections for well-represented groups of rats (*Sundamys* and *Rattus*) as models to evaluate the interplay of Plio-Pleistocene changes in the diversification patterns of mammals in Sundaland. I use genetics in combination with morphology and ecology to look at diversification both within and between species in these two related genera to test biogeographic and other evolutionary hypotheses.

Chapter 1, Introduction. I describe the theoretical framework of my study, introduce the biological study models, set the hypotheses underlying my research and overview the methodological approach to tackle these questions, including methodological challenges. The main outcomes of this chapter are two publications:

1. **Camacho-Sanchez, M**., Burraco, P., Gomez-Mestre, I., & Leonard, J. A. (2013). Preservation of RNA and DNA from mammal samples under field conditions. **Molecular Ecology Resources,** 13(4), 663-673 (Appendix 1.1).
2. Brandariz-Fontes, C*., **Camacho-Sanchez, M***., Vilà, C., Vega-Pla, J. L., Rico, C., & Leonard, J. A. (2015). Effect of the enzyme and PCR conditions on the quality of high-throughput DNA sequencing results. **Scientific Reports**, 5, 8056 (*equal contribution; Appendix 1.2).

Chapter 2, Interglacial refugia on tropical mountains: novel insights from the summit rat (*Rattus baluensis*), a Borneo mountain endemic. I study the role of tropical mountains as interglacial refugia. I use genetics to evaluate the effect of post-Last Glacial Maximum (~21 Kya) changes on the demographic history of *Rattus baluensis*, a rat endemic to habitats above 2000 m in northern Borneo.

Chapter 3, Rapid external morphological divergence after mountain colonization in a Sunda rat. I discuss the phylogenetic relationships of Sundaic endemic *Rattus* with particular focus to the evolution of the montane lineages.

Chapter 4, The generic status of *Rattus annandalei* Bonhote, 1903 (Rodentia, Murinae) and its evolutionary implications. I generate a complete phylogeny of the genus *Sundamys*, reclassify *Rattus annandalei* as *Sundamys annandalei*, and discuss its evolutionary implications.

Chapter 5, Multilocus nuclear and mitogenome DNA analyses expose complex genetic structure in *Sundamys* rats across Sundaland. I construct a comprehensive phylogeographic study of *Sundamys*, a rat genus endemic to Sundaland, as a model to evaluate Quaternary changes on vertebrate diversification in this region.

Resumen

La bioregión tropical de Sunda, Sudeste asiático, es uno de los principales puntos calientes de biodiversidad mundial, incluyendo la de mamíferos. Su compleja historia geológica proporciona un excelente escenario para estudiar evolución. Sin embargo, el estudio de la biogeografía de este grupo en Sunda ha estado limitado por (1) la incertidumbre taxonómica para gran parte de sus grupos, y (2) por el muestreo sesgado o falta de datos en muchas zonas (por ejemplo, Sumatra). En esta tesis combino un enfoque taxonómico acompañado con un amplio muestreo de campo y en colecciones de historia natural para dos grupos de ratas, *Sundamys* y *Rattus*, como modelos para evaluar la interacción de cambios en el Plio-Pleistoceno con la diversificación de mamíferos en Sunda. Empleo herramientas genéticas combinadas con morfología y ecología para estudiar los patrones de diversificación intra/inter específicos en estos dos géneros para testar hipótesis biogeográficas y evolutivas.

Capítulo 1, Introducción. Describo el marco teórico de mi estudio, introduzco los modelos biológicos, establezco las hipótesis que sustentan mi investigación y proporciono una visión en conjunto de los métodos desarrollados para abordarla. El principal resultado de este capítulo son dos publicaciones:

1. **Camacho-Sanchez, M**., Burraco, P., Gomez-Mestre, I., & Leonard, J. A. (2013). Preservation of RNA and DNA from mammal samples under field conditions. **Molecular Ecology Resources,** 13(4), 663-673 (Appendix 1.1).
2. Brandariz-Fontes, C*., **Camacho-Sanchez, M***., Vilà, C., Vega-Pla, J. L., Rico, C., & Leonard, J. A. (2015). Effect of the enzyme and PCR conditions on the quality of high-throughput DNA sequencing results. **Scientific Reports**, 5, 8056 (*equal contribution; Appendix 1.2).

Capítulo 2, Refugios interglaciares en montañas tropicales: nuevas perspectivas a partir del estudio de *Rattus baluensis*, una rata endémica de zonas montañosas. Estudio el papel de las montañas tropicales como refugios interglaciares. Uso la genética para evaluar el efecto de cambios desde del Último Máximo Glacial, aproximadamente 21 mil años atrás, en la demografía histórica de *Rattus baluensis*, una rata endémica de zonas montañosas superiores a 2,000 m en el norte de Borneo.

Capítulo 3, Rápida divergencia en la morfología externa una rata de Sunda tras la colonización de zonas montañosas. Discuto las relaciones filogenéticas de especies de *Rattus* endémicas a Sunda, con especial énfasis en le evolución de linajes de montaña.

Capítulo 4, Revisión taxonómica de *Rattus annandalei* Bonhote, 1903 (Rodentia, Murinae) y sus implicaciones evolutivas. Genero una filogenia completa para el género *Sundamys*, reclasifico *Rattus annandalei* como *Sundamys annandalei*, y discuto las implicaciones evolutivas derivadas.

Capítulo 5, Análisis con mitogenomas y marcadores nucleares exponen la compleja diversidad genética de *Sundamys* en Sundaland. Reconstruyo una filogeografía completa para *Sundamys*, un género de ratas endémico de Sunda, como modelo para evaluar los cambios del Cuaternario en la diversificación de vertebrados en esta región.

Table of Contents

Summary I
Resumen III
Table of Contents V
List of Figures VII
List of Tables VIII
Chapter 1 Introduction 1
Sundaland 1
Biogeographical definition and geological history of Sundaland 2
Goals and hypotheses 4
Study system 4
Sampling 4
Molecular markers 7
Literature cited 10
Appendix 1: PUBLICATION. Preservation of RNA and DNA from mammal samples under field conditions 13
Appendix 2: PUBLICATION. Effect of the enzyme and PCR conditions of the quality of high-throughput DNA sequencing results 24
Chapter 2 Interglacial refugia on tropical mountains: novel insights from the summit rat (*Rattus baluensis*), a Borneo mountain endemic 29
Introduction 31
Methods 32
Results 41
Discussion 47
Acknowledgements 51
Literature cited 51
Appendix 2.1. Amplicon library preparation of intron markers 56
Appendix 2.2. PCA of the rejection step by PopABC 61
Appendix 2.3. *Nepenthes rajah* and summit rat catches on Tambuyukon 62
Appendix 2.4. Evanno Delta K 63
Appendix 2.5. K2 to K6 from STRUCTURE 64
Appendix 2.6. *cytb* and control region haplotype networks 65
Appendix 2.7. Connectivity reconstructions 66
Chapter 3 Rapid external morphological divergence after mountain colonization in a Sunda rat 68
Introduction 70
Methods 73
Results 79
Discussion 82
Acknowledgements 86
Literature cited 87

Appendix 3.1 Haplotype assignation 91
Chapter 4 The generic status of *Rattus annandalei* Bonhote, 1903 (Rodentia, Murinae) and its evolutionary implications 92
Introduction 94
Materials and Methods 95
Results 103
Discussion 113
Acknowledgements 115
Literature Cited 117
Appendix 4.1. Illumina library preparation 122
Appendix 4.2. Specimens used for the palatal view of the skull in geometric morphometric analysis 127
Appendix 4.3. Specimens used for the dentary lateral side in the morphometric geometric analysis 128
Appendix 4.4. Per-locus PhyloBayes trees. 129
Chapter 5 Multilocus nuclear and mitogenome DNA analyses expose complex genetic structure in *Sundamys* rats across Sundaland 130
Introduction 132
Methods 133
Results 147
Discussion 152
Acknowledgements 155
Literature cited 156
Appendix 5.1. Sequencing data. 160
Appendix 5.2. RAxML maximum likelihood tree with protein-coding genes of mitogenomes (all samples shown). 164
Appendix 5.3. TCS haplotype networks for the nuclear loci. 165
Conclusions 167
Acknowledgements 169

List of Figures

Figure 1-1. Map of Sundaland. 3
Figure 1-2. Field work localities. 6
Figure 1-3. Sequencing yield for introns from multiplex PCRs. 8
Figure 1-4. Average coverage and proportion mitogenomes reconstructed. 8
Figure 1-5. Field work in northern Borneo. 9
Figure 2-1. A) Study area. B) Summit rat licking nectar from *Nepenthes rajah* and mountain scrubland habitat. C) Elevation profile across Mt. Kinabalu and Mt. Tambuyukon, with trapping effort D) Ancestry from STRUCTURE. 35
Figure 2-2. TCS haplotype networks for 19 polymorphic introns. 44
Figure 2-3. TCS haplotype network for the mitogenomes. 45
Figure 2-4. PopABC demographic reconstructions 47
Figure 3-1. Distribution of Sunda native *Rattus* 71
Figure 3-2. Skins of mountain vs lowland *Rattus* 72
Figure 3-3. RAxML consensus tree of mitogenomes 80
Figure 3-4. Haplotype network of *cyt b* from *R. baluensis* and *R. tiomanicus* 82
Figure 3-5. Dated tree from BEAST inference in *Rattus*. 84
Figure 4-1. Distribution of *Rattus annandalei* and *Sundamys*. 95
Figure 4-2. Landmark locations and definitions 102
Figure 4-3. PhyloBayes tree from mitogenomes in *Sundamys* 104
Figure 4-4. Dated tree from BEAST inference in *Sundamys* 105
Figure 4-5. PCA and 3-way discriminant analyses of cranium among *Sundamys* and Indo-Pacific *Rattus* 106
Figure 4-6. PCA of the cranium and mandible *Sundamys* - *Rattus annandalei*. 108
Figure 4-7. Skull and jaw views in *Rattus* and *Sundamys*. 112
Figure 5-1. Distribution and sampling of *Sundamys* 133
Figure 5-2. RAxML tree from *Sundamys* mitogenomes 144
Figure 5-3. MDS of the nuclear variation in the *muelleri* group 147
Figure 5-4. *cyt b* haplotype network of *Sundamys muelleri* 148
Figure 5-5. Pairwise *cytb* uncorrected genetic distances and genotype Prevosti's distance for lineages within *Sundamys* 150
Figure 5-6. *Sundamys* multilocus species trees. 152

List of Tables

Table 2-1. Samples included in the study of *Rattus baluensis*. 36

Table 2-2. Primers used to amplify the 27 introns in *R. baluensis* 37

Table 2-4. Prior and posterior distributions of PopABC simulations 41

Table 2-4. Nuclear genetics statictics on *R. baluensis*. 43

Table 2-5. Mitochondrial genetic diversity of *Rattus baluensis*. 46

Table 3-1. Field samples and museum specimens sequenced of *Rattus*. 75

Table 3-2. Sequencing information of *Rattus* mitogenomes. 80

Table 4-1. *Rattus* sequences used for phylogenetic reconstructions 97

Table 4-2. Genetic distances between *R. annandalei* and *Sundamys* 105

Table 4-3. Selected external measurements of adult *Sundamys* 109

Table 5-1. Distrubution of *Sundamys* and sampling scheme. 135

Table 5-2. Pairwise genetic distances of *cytb* between *Sundamys* lineages 151

Chapter 1 Introduction

Sundaland

Sundaland is a biogeographic region in tropical Southeast Asia (Figure 1.1) and a leading world hotshot of biodiversity. For instance, 39 % of the vertebrates are endemic and it hosts 5 % of worldwide plant endemism (Myers 2000). The unique biodiversity of this region inspired Wallace to independently arrive at the theory of evolution by natural selection, and has played a prominent role in early development of the field of biogeography.

Research in taxonomy (e.g. Musser and Califia 1982; Musser and Newcomb 1983) and biogeography (e.g. Heaney 1978, 1984, 1986), have recurrently brought up the topic of the complex geology of Sundaland has driver of the biodiversity patterns we can observe nowadays. The recent incorporation of genetics is providing solid basis for addressing taxonomic issues and unveiling a great cryptic diversity for some mammals from this region (e.g. Ruedi and Fumagalli 1996; den Tex et al. 2010; Esselstyn et al. 2013; Demos et al. 2016; Hawkins et al. 2016; Mason et al. 2016), while phylogeographic patterns and their connections with the geological history start to arise (reviews in Lohman et al. 2011; de Bruyn et al. 2014; Leonard et al. 2015; Sheldon et al. 2015). As a result, mountain driven diversity, and divergence in allopatry caused by sea level oscillations and ecological barriers (vegetation changes) across the turbulent Plio-Pleistocene, have led to patterns like mountain endemics (den Tex et al. 2010; Esselstyn et al. 2013; Demos et al. 2016), the relative isolation of Java (van der Bergh 2001), the proximity of Sumatran and Peninsular lineages (Leonard et al. 2015), or the great diversity and divergence for some birds within Borneo (Sheldon et al. 2015). However, most of the vertebrates have been understudied and their taxonomy is uncertain. Furthermore, much of the studies are restricted to a single taxon, a single island or lack the genetic power to address specific hypothesis.

This dissertation aims to assess the effects of Plio-Pleistocene changes in vertebrate diversification and evolution in Sundaland, using multilocus genetics in a comparative phylogeographic framework with two groups of understudied Sunda rats.

Biogeographical definition and geological history of Sundaland

Wallace's Line separates Sundaland from Wallacea to the southeast (Figure 1.1). Although the distance between Bali and Lombok and Borneo and Sulawesi is small, the sea between them is deep, and these islands have never been connected by land. A northwards branch of Wallace's Line called Huxley's Line sets a biogeographical break between Sundaland and the Philippines to the east. Palawan is a northeast extension of the Sunda Shelf and is now regarded as a transition zone, referred to as the Huxley Filter Zone (Esselstyn et al. 2010). The northern limit of Sundaland is the Isthmus of Kra on the Thai-Malay Peninsula. It is a transition zone between the Sunda and Indo-Burma bioregions (Parnell 2013). At its narrowest point (10º 11N) the isthmus narrows to only 45 km and the land reaches up to 75 meters above sea level (Parnell 2013). Although some studies hypothesize a possible water gateway could have prevailed during long periods and favored this transition (de Bruyn et al. 2005), there is no clear evidence for such a water barrier. On the contrary, a strong climatic gradient, currently present, seems to have persisted during past glacials and interglacials at the latitude of the Isthmus of Kra (Hughes et al. 2011). Attempts have been made to explain patterns of species distributions of mammals across the Isthmus by repeated sea level rises (>58 times above 40 m) during the last 5 My which, could also have led to compression and extinction of populations around the Isthmus (Woodruff and Turner 2009). Perhaps the strong climatic gradient acts combined with the narrowing around the Isthmus (acerbated during episodes of sea-level rise) for enhancing this biogeographical barrier.

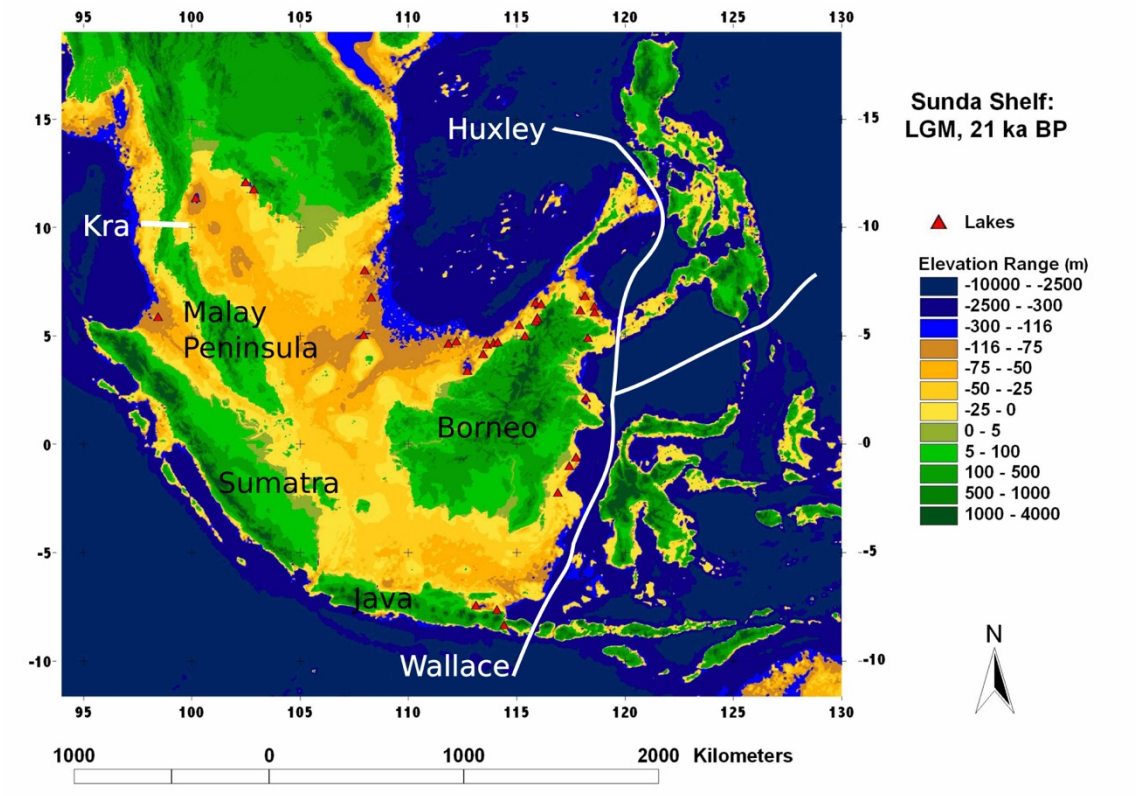


Figure 1-1. Map of Sundaland and its biogeographical limits (base map from Sathiamurthy and Voris 2006; biogeographical lines from Esselstyn et al. 2010).

Sundaland extends across the Sunda Shelf, which is south-easternmost extension of the Eurasian Plate. Its geologic origin dates back from the Early Mesozoic (252 Mya), although its current configuration is from the Pliocene (Hall 2002, 2009). Whereas the tectonic dynamics during the Cenozoic (66 Mya to present) have been very complex for the surrounding areas of the Philippines and Wallacea, as the Australian Plate moved northwards, the conditions of the Sunda Shelf have been very stable (Hall 2002, 2009). However, sea level has fluctuated regularly since the late Pliocene (Miller et al. 2005). These large fluctuations exposed large areas of the Sunda Shelf during glacial maxima, connecting many islands with each other and mainland. Although large areas of the Shelf were exposed and connected lands which are currently isolated, the habitat on that land is not completely characterized. Past habitat modeling suggests that both highland and lowland forests were more extensive during glacial maxima (Cannon et al. 2009). That does not exclude the possibility of large dry vegetation blocks crossing between Borneo and Sumatra (Bird et al. 2005; Cannon et al. 2009).

Goals and hypotheses

The underlying working hypotheses which support my study are:

1. The genetic diversity in Sunda mammals has been shaped by sea level oscillations and vegetation changes derived from the profound climatic oscillations in Plio-Pleistocene. Much of this genetic diversity could be very divergent and cryptic in lineages from different islands.
2. Mountains promote diversity in Sunda as (1) reservoirs of genetic diversity providing refugia during climatic oscillations and (2) by rendering a different ecological matrix to lowland forest to which lineages could have adapted and ultimately diverged.

Study system

Small mammals of the Rattini (Order Rodentia) were selected as models to test the evolutionary hypothesis. They offer several advantages due both to their evolutionary history, rate of speciation, ecological specialization as well as practical advantages such as getting permits, trapability and relatedness to a model organism.

Several genera of rats have their center of diversification in Sundaland (Musser and Newcomb 1983). Two of those genera are *Sundamys* and *Rattus.* The genus *Sundamys* is endemic to Sundaland, whereas the genus *Rattus* is much more widespread. Both genera have several lowland species (*Rattus tiomanicus*, *Sundamys muelleri* and *S. annandalei*) and mountain endemics (*Sundamys infraluteus*, *S. maxi*, *Rattus baluensis*, *R. korinchi*, and *R. hoogerwerfi*; Musser and Newcomb 1983; Musser 1986; Musser and Carleton 2005). The time of diversification for both genera seems to correlate with the Plio-Pleistocene, the geological period we are interested in (Aplin et al 2011; Fabre et al. 2013).

Furthermore, I profited from their close evolutionary affinity with *Rattus norvegicus*, a model species in biomedical research for which there are genomic resources, including a well-annotated genome for marker development (rgd.mcw.edu; Shimoyama et al. 2015).

Sampling

I targeted sampling all Sundaland native species in the two target genera and all main populations of the widespread species. This was achieved by sampling fresh tissue

during three fieldwork campaigns in Borneo, which was complimented through exchange and loans from collaborators, and samples from public repositories such as tissue collections and historical specimens in museums.

Field work in Borneo

Due to the remote location of some field localities, we tested a buffer to preserve mammalian tissue samples under field conditions for genetic (both DNA and RNA) analysis. We published the results in *Molecular Ecology Resources* (Camacho-Sanchez et al. 2013; Appendix 1.1).

We undertook three fieldwork campaigns in Sabah, northern Borneo (Figure 1.5): (1) June-August 2012 to Mount (Mt.) Tambuyukon in Kinabalu National Park; (2) February-April 2013 to Mt. Tambuyukon and Mt. Kinabalu (Hawkins 2015); and (3) June-July 2016 to Mt. Alab, Cocker Range, and Mt. Trusmadi (Figure 1.2). The goals of this fieldwork were to generate a collection of small mammal samples for genetics, ecology and taxonomy, and to describe distribution patterns. We developed a trapping strategy to maximize our trapping success in terms of number of catches for abundant ground-dwelling small mammals combined with diversity by deploying parallel trapping strategies which yielded abundant samples for most common species, as well as few samples from the more specialized mammals, such as shrews or arboreal mammals. We mainly targeted rats, treeshrews and squirrels. We used wire cage Tomahawk traps and Sherman traps of different sizes, and a combined bait which could include banana, vanilla scent, palm fruit, sweet potato, coconut and dry fish. We were able to have a broad representation of the ground small mammal community, as well as some arboreal species.

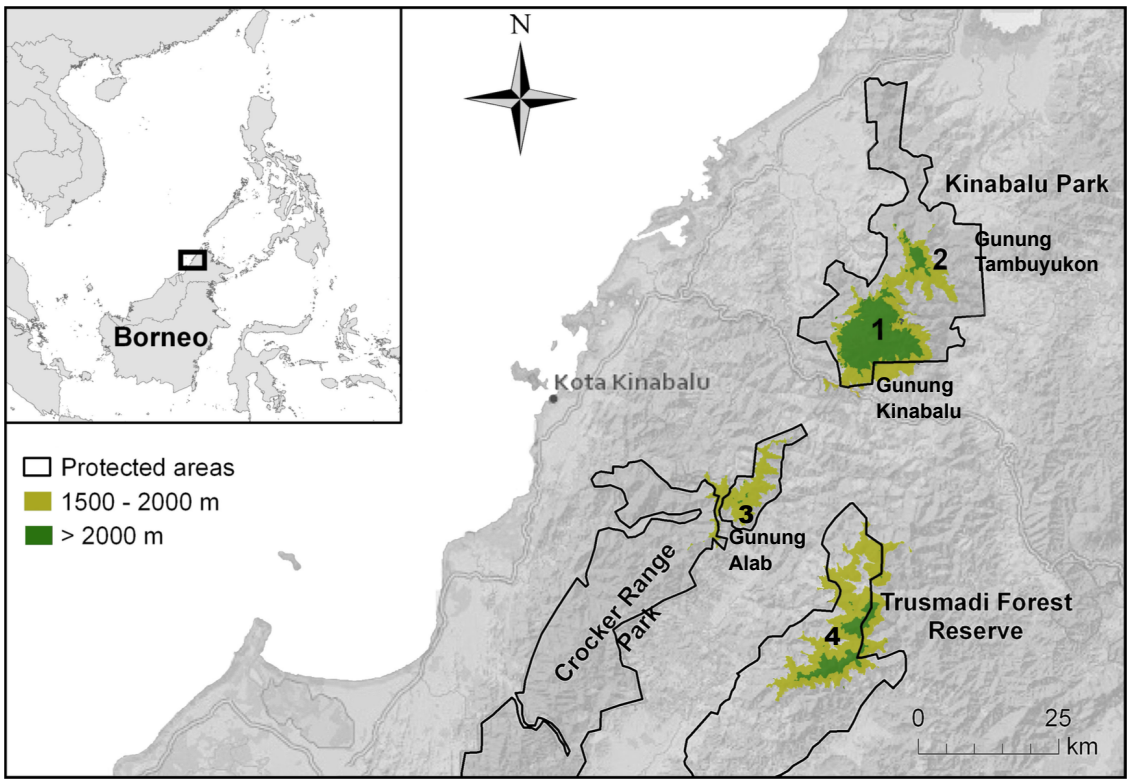


Figure 1-2. Field work localities (map adapted from Quintanilla 2014).

The first expedition targeted Mt. Tambuyukon, the third highest peak in Borneo (2,579 m). It represents the first elevational survey for mammals in this mountain. We trapped at lowland forest starting at Monggis substation 350 m, moved the camp up to Kepuakan Camp (900), sampled mountain forest in Musang Camp (1430 m), mossy forest in Jeneral Camp (2,030 m) and the summit forest from Rajah Camp (2,400 m). In the second trip, we surveyed Mt. Kinabalu from the Park Headquarters up to around Waras, Pendant hut and Panar Laban (3,200 m), a resurvey in the last 500 m in Tambuyucon, from Jeneral Camp to the summit, and in Poring Hot Springs, from the entrance (500 m) and along the Langanan Trail up to the Langanan waterfall (900 m). Details in Hawkins (2015). In the final expedition, we sampled at the summit area of Mt. Alab (1800-1900 m), a peak in Crocker Range Park; and Mt. Trusmadi, the 2nd highest peak in Borneo, from 900 m, in Mirad Irad Camp, to the summit (2,642 m; Figure 1.5).

Historical samples from museum specimens

We requested historic material whenever fresh samples were not obtainable for the given taxon. Damage to historical specimens was minimized by focusing the sampling

to specific taxonomic units or population of special interest for which no modern material was available, and by sampling preferentially dry tissue attached to the skulls or clips from skins whenever skulls were clean. These specimens had been collected along the 20th century. Natural history collections in North America, Europe and Asia were visited.

Molecular markers

Many phylogeographic and phylogenetic studies of small mammals in Sundaland and elsewhere have been undertaken in the last couple of decades (review in Leonard 2015; Esselstyn et al. 2013; Demos et al. 2016). By far the molecular marker of choice in these studies is the mitochondrial cytochrome *b*. This marker can be amplified with universal primers, and is generally variable and informative within species. Although it is useful for looking at patterns within species, it can be insufficient to resolve nodes in phylogenetic reconstructions. For this reason, we targeted the whole mitochondrial, which has been shown to be more informative at different evolutionary scales (e.g. Robins et al. 2008). Mitochondrial markers, either just cytochrome *b* or the whole mitochondrial genome, is a single genetic marker with a low (compared to nuclear DNA) effective population size, making is more subject to drift and introgression (Zink and Barrowclough 2008; Hailer et al. 2012; Pagès et al. 2013). For these reasons, including nuclear markers that sort independently is important for elucidating species relationships and histories (Edwards 2009). We targeted a small panel of nuclear markers which is commonly used in rodents thus could be compared directly to more information from previous studies (~4500 bp), and also developed larger panel of nuclear markers selected from Igea et al. (2010), which were variable both within and between these closely related animals (~12000 bp; Methods in Chapter 2 for details). We developed protocols for amplification of the intron panel in multiplexed PCRs (see Methods in Chapter 2), which included using high-fidelity DNA polymerases and reduced cycling conditions (see our publication in Scientific Reports, Brandariz-Fontes et al. 2015, in Appendix 1.2). They provided a good lab-work investment-yield trade-off (Figure 1.3).

Different strategies were developed to sequence these different loci. Most nuclear loci were polymorphic at the population scale (Chapter 2). Given their sequence-based nature we further used them across several species for multilocus phylogenetic inference (Chapter 5; Figure 5.6).

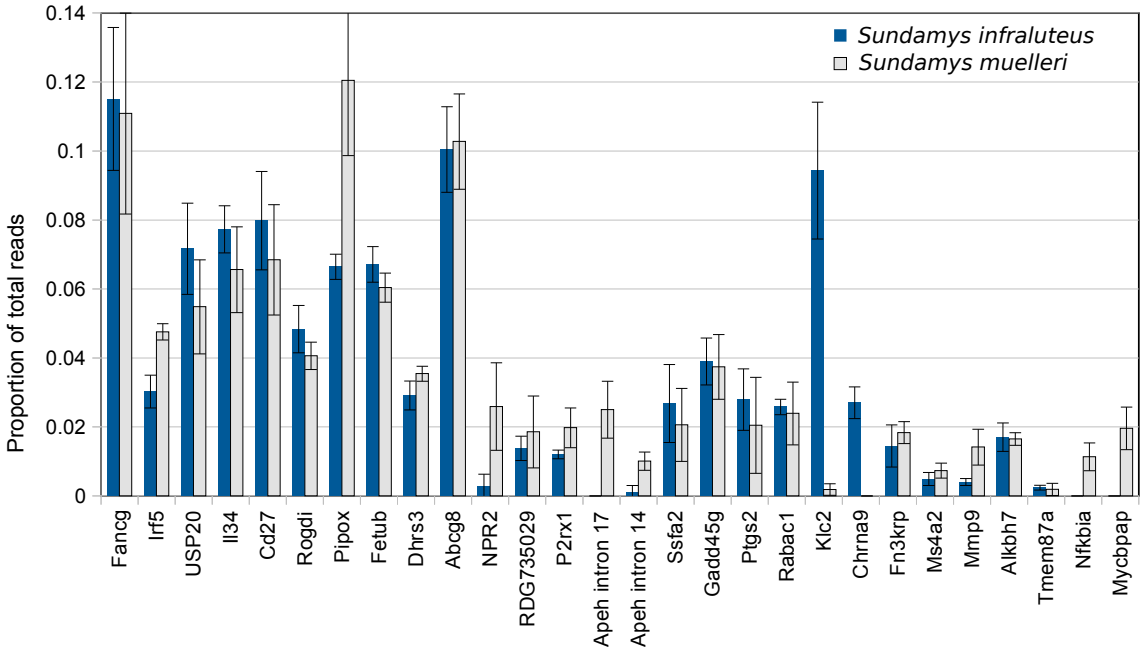


Figure 1-3. Example of relative sequencing yield for the introns amplified in multiplex PCRs for 3 random samples in *S. muelleri* and *S. infraluteus*.

The different strategies we used for sequencing mitogenomes (Methods in Chapters 2-5) yielded varying outputs. An average coverage of 10x was often enough to assemble mitochondrial genomes (Fig 1.4).

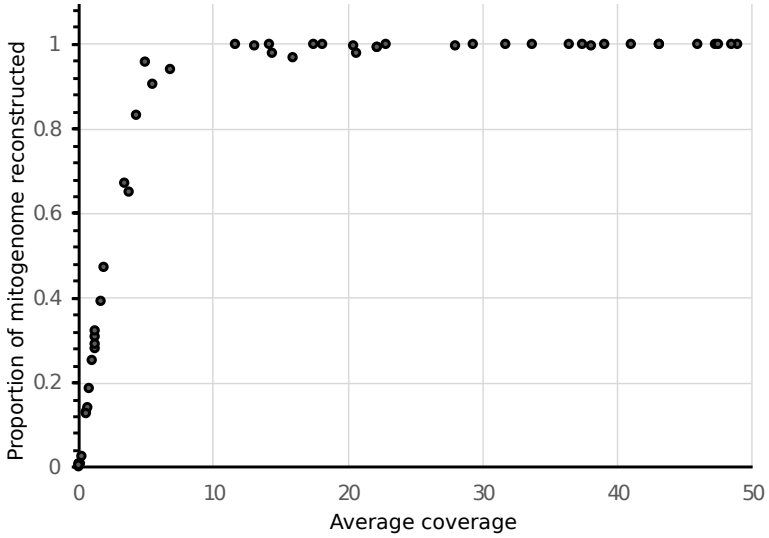


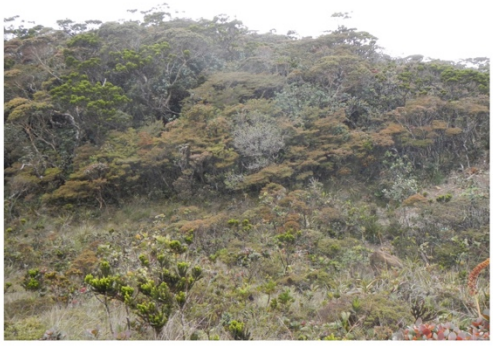
Figure 1-4. Relation between average coverage and proportion of mitogenome reconstructed (based on data from Appendix 5.2).

Preparing traps in Mirad Irad Camp, Trusmadi

Views from summit trapping locations in Trusmadi

Tomahawk trap

Clouds covering the camp at 2000 m in Mt. Tambuyukon



Mountain scrubland at 2500 m in Mt. Tambuyukon

Mossy forest at 2000 m in Mt. Tambuyukon

Figure 1-5. Field work in northern Borneo.

Literature cited

APLIN, K. P. ET AL. 2011. Multiple Geographic Origins of Commensalism and Complex Dispersal History of Black Rats. PLoS ONE 6:e26357.

BRANDARIZ-FONTES, C., M. CAMACHO-SANCHEZ, C. VILÀ, J. L. VEGA-PLA, C. RICO AND J. A. LEONARD. 2015. Effect of the enzyme and PCR conditions on the quality of high-throughput DNA sequencing results. Scientific Reports 5:8056.

CAMACHO-SANCHEZ, M., P. BURRACO, I. GOMEZ-MESTRE AND J. A. LEONARD. 2013. Preservation of RNA and DNA from mammal samples under field conditions. Molecular Ecology Resources 13:663–673.

DE BRUYN, M., E. NUGROHO, M. M. HOSSAIN, J. C. WILSON AND P. B. MATHER. 2005. Phylogeographic evidence for the existence of an ancient biogeographic barrier: the Isthmus of Kra Seaway. Heredity 94:370–378.

DE BRUYN, M. ET AL. 2014. Borneo and Indochina are Major Evolutionary Hotspots for Southeast Asian Biodiversity. Systematic Biology 63:879–901.

DEMOS, T. C. ET AL. 2016. Local endemism and within-island diversification of shrews illustrate the importance of speciation in building Sundaland mammal diversity. Molecular Ecology 25:5158–5173.

DEN TEX, R.-J., R. THORINGTON, J. E. MALDONADO AND J. A LEONARD. 2010. Speciation dynamics in the SE Asian tropics: Putting a time perspective on the phylogeny and biogeography of Sundaland tree squirrels, *Sundasciurus*. Molecular Phylogenetics and Evolution 55:711–20.

EDWARDS, S. V. 2009. Is a new and general theory of molecular systematics emerging? Evolution 63:1–19.

ESSELSTYN, J. A, MAHARADATUNKAMSI, A. S. ACHMADI, C. D. SILER AND B. J. EVANS. 2013. Carving out turf in a biodiversity hotspot: multiple, previously unrecognized shrew species co-occur on Java Island, Indonesia. Molecular Ecology 22:4972–4987.

ESSELSTYN, J. A., C. H. OLIVEROS, R. G. MOYLE, A. T. PETERSON, J. A. MCGUIRE AND R. M. BROWN. 2010. Integrating phylogenetic and taxonomic evidence illuminates complex biogeographic patterns along Huxley's modification of Wallace's Line. Journal of Biogeography 37:2054–2066.

FABRE, P. ET AL. 2013. A new genus of rodent from Wallacea (Rodentia: Muridae: Murinae: Rattini), and its implication for biogeography and Indo-Pacific Rattini systematics. Zoological Journal of the Linnean Society 169:408–447.

HAILER, F. ET AL. 2012. Nuclear genomic sequences reveal that polar bears are an old and distinct bear lineage. Science 336:344–347.

HALL, R. 2009. Southeast Asia's changing palaeogeography. Blumea 54:148–161.

HALL, R. 2002. Cenozoic geological and plate tectonic evolution of SE Asia and the SW Pacific: computer-based reconstructions, model and animations. Journal of Asian Earth Sciences 20.

HAWKINS, M. T. R. 2015. Biogeography and Phylogeography of Mammals of Southeast Asia: A Comparative Analysis Utilizing Macro and Microevolution. Doctoral thesis. George Mason University.

HAWKINS, M. T. R., K. M. HELGEN, J. E. MALDONADO, L. L. ROCKWOOD, M. T. N. TSUCHIYA AND J. A. LEONARD. 2016. Phylogeny, biogeography and systematic revision of plain long-

nosed squirrels (genus *Dremomys*, Nannosciurinae). Molecular Phylogenetics and Evolution 94:752–764.

HEANEY, L. R. 1986. Biogeography of mammals in SE Asia: estimates of rates of colonization, extinction and speciation. Biological Journal of the Linnean Society 28:127–165.

HEANEY, L. R. 1978. Island Area and Body Size of Insular Mammals: Evidence from the Tri-Colored Squirrel (*Callosciurus prevostii*) of Southeast Asia. Evolution 32:29–44.

HEANEY, L. R. 1984. Mammalian species richness on islands on the Sunda Shelf, Southeast Asia. Oecologia 61:11–17.

HUGHES, A. C., C. SATASOOK, P. J. J. BATES, S. BUMRUNGSRI AND G. JONES. 2011. Explaining the causes of the zoogeographic transition around the Isthmus of Kra: Using bats as a case study. Journal of Biogeography 38:2362–2372.

HUGHES, J. B., P. D. ROUND AND D. S. WOODRUFF. 2003. The Indochinese-Sundaic faunal transition at the Isthmus of Kra: an analysis of resident forest bird species distributions. Journal of Biogeography 30:569–580.

IGEA, J., J. JUSTE AND J. CASTRESANA. 2010. Novel intron markers to study the phylogeny of closely related mammalian species. BMC Evolutionary Biology 10:369.

LEONARD, J. A., R. J. DEN TEX, M. T. R. HAWKINS, V. MUÑOZ-FUENTES, R. THORINGTON AND J. E. MALDONADO. 2015. Phylogeography of vertebrates on the Sunda Shelf: A multi-species comparison. Journal of Biogeography 42:871–879.

LOHMAN, D. J. ET AL. 2011. Biogeography of the Indo-Australian Archipelago. Annual Review of Ecology, Evolution, and Systematics 42:205–226.

MASON, V. C. ET AL. 2016. Genomic analysis reveals hidden biodiversity within colugos, the sister group to primates. Science Advances 2:e1600633–e1600633.

MILLER, K. G. 2005. The Phanerozoic Record of Global Sea-Level Change. Science 310:1293–1298.

MUSSER, G. G. AND D. CALIFIA. 1982. Identities of rats from Pulau Maratua and other islands off East Borneo. American Museum Novitates:1–30.

MUSSER, G. G. 1986. Sundaic *Rattus*: definitions of *Rattus baluensis* and *Rattus korinchi*. American Museum Novitates 2862:1–24.

MUSSER, G. G. AND M. D. CARLETON. 2005. Superfamily Muroidea. Pp. 894–1531 in Mammal Species of the World: a taxonomic and geographic reference (D. E. Wilson & D. M. Reeder, eds.). 3rd edition. The Johns Hopkins University Press, Baltimore.

MUSSER, G. G. AND C. NEWCOMB. 1983. Malaysian murids and the giant rat from Sumatra. Bulletin of the American Museum of Natural History 174:327–598.

MYERS, N., R. A. MITTERMEIER, C. G. MITTERMEIER, G. A. B. DA FONSECA AND J. KENT. 2000. Biodiversity hotspots for conservation priorities. Nature 403:853–8.

PAGÈS, M. ET AL. 2013. Cytonuclear discordance among Southeast Asian black rats (*Rattus rattus* complex). Molecular Ecology 22:1019–34.

PARNELL, J. 2013. The biogeography of the Isthmus of Kra region: A review. Nordic Journal of Botany 31:001–015.

QUINTANILLA, I. 2014. A novel genomic approach to study the phylogeography of a rat, *Rattus baluensis*, using intron sequences and complete mitochondrial genomes. University Pablo de Olavide.

RUEDI, M. AND L. FUMAGALLI. 1996. Genetic structure of Gymnures (genus *Hylomys*; Erinaceidae) on continental islands of Southeast Asia: historical effects of fragmentation. Journal of Zoological Systematics and Evolutionary Research 34:153–162.

SATHIAMURTHY, E. AND K. H. VORIS. 2006. Maps of Holocene Sea Level Transgression and Submerged Lakes on the Sunda Shelf. P. in The Natural History Journal of Chulalongkorn University.

SHELDON, F. H., H. C. LIM AND R. G. MOYLE. 2015. Return to the Malay Archipelago: the biogeography of Sundaic rainforest birds. Journal of Ornithology. doi:10.1007/s10336-015-1188-3

SHIMOYAMA, M. ET AL. 2015. The Rat Genome Database 2015: Genomic, phenotypic and environmental variations and disease. Nucleic Acids Research 43:D743–D750.

VAN DEN BERGH, G. D., J. DE VOS AND P. Y. SONDAAR. 2001. The Late Quaternary palaeogeography of mammal evolution in the Indonesian Archipelago. Palaeogeography, Palaeoclimatology, Palaeoecology 171:385–408.

WOODRUFF, D. S. AND L. M. TURNER. 2009. The Indochinese-Sundaic zoogeographic transition: a description and analysis of terrestrial mammal species distributions. Journal of Biogeography 36:803–821.

ZINK, R. M. AND G. F. BARROWCLOUGH. 2008. Mitochondrial DNA under siege in avian phylogeography. Molecular Ecology 17:2107–2121.

Appendix 1: PUBLICATION. Preservation of RNA and DNA from mammal samples under field conditions

MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2013) doi: 10.1111/1755-0998.12108

Preservation of RNA and DNA from mammal samples under field conditions

MIGUEL CAMACHO-SANCHEZ,* PABLO BURRACO,† IVAN GOMEZ-MESTRE† and JENNIFER A. LEONARD*

*Conservation and Evolutionary Genetics Group, Estación Biológica de Doñana (EBD-CSIC), C/Américo Vespucio, s/n, 41092 Seville, Spain, †Ecology, Evolution, and Development Group, Estación Biológica de Doñana (EBD-CSIC), C/Américo Vespucio, s/n, 41092 Seville, Spain

Abstract

Ecological and conservation genetics require sampling of organisms in the wild. Appropriate preservation of the collected samples, usually by cryostorage, is key to the quality of the genetic data obtained. Nevertheless, cryopreservation in the field to ensure RNA and DNA stability is not always possible. We compared several nucleic acid preservation solutions appropriate for field sampling and tested them on rat (*Rattus rattus*) blood, ear and tail tip, liver, brain and muscle. We compared the efficacy of a nucleic acid preservation (NAP) buffer for DNA preservation against 95% ethanol and Longmire buffer, and for RNA preservation against RNA*later* (Qiagen) and Longmire buffer, under simulated field conditions. For DNA, the NAP buffer was slightly better than cryopreservation or 95% ethanol, but high molecular weight DNA was preserved in all conditions. The NAP buffer preserved RNA as well as RNA*later*. Liver yielded the best RNA and DNA quantity and quality; thus, liver should be the tissue preferentially collected from euthanized animals. We also show that DNA persists in nonpreserved muscle tissue for at least 1 week at ambient temperature, although degradation is noticeable in a matter of hours. When cryopreservation is not possible, the NAP buffer is an economical alternative for RNA preservation at ambient temperature for at least 2 months and DNA preservation for at least 10 months.

Keywords: degradation, field sampling, NAP buffer, RNA*later*, tissue storage

Received 31 October 2012; revision received 18 March 2013; accepted 21 March 2013

Introduction

It is critical to all genetic studies based on field samples to preserve them properly from point of collection to the laboratory. Good preservation of samples that may be used for genomic studies is even more important because many genomic protocols require a high quantity of high-quality nucleic acids (Wong *et al.* 2012). Genomic techniques such as next-generation sequencing are becoming increasingly popular because they have allowed researchers to expand from transcriptome and genome experiments on model organisms in the laboratory, to applying these tools to specific ecological and evolutionary questions in nonmodel organisms in the wild (Dassanayake *et al.* 2009; Elmer *et al.* 2010; Hohenlohe *et al.* 2010; Wolf *et al.* 2010; Chen *et al.* 2011). However, many interesting biological samples for molecular ecology occur in locations where their preservation for genetic and expression studies is difficult, and cryopreservation is not possible. In this context, preservation of high-quantity and high-quality DNA and RNA under field conditions is fundamental to many new molecular ecology studies.

DNA and RNA degrade with increased time and temperature (Ludes *et al.* 1993; Vincek *et al.* 2003; Seear & Sweeney 2008), and RNA degrades more rapidly than DNA (Massie *et al.* 1972). The best way to preserve RNA is to snap-freeze samples in liquid nitrogen followed by storage at −80 °C (Gorokhova 2005; Wang & Sherman 2006; Riesgo *et al.* 2012). However, cryopreservation in the field can be difficult or impossible. Stabilizing buffers such as RNA*later* (Qiagen) can preserve RNA at ambient temperature (Vincek *et al.* 2003; Gorokhova 2005; Gayral *et al.* 2011). However, they are expensive and fieldwork often extends beyond time and/or temperature conditions suggested by the manufacturers (i.e. RNA*later* is approved for storing tissue samples 4 weeks at 2–8 °C, up to 7 days at 15–25 °C or up to 1 day at 37 °C).

Correspondence: Miguel Camacho-Sanchez, Fax: (+34) 954621125; E-mail: miguel.camacho@ebd.csic.es

Cryopreservation is also the best way to preserve DNA (Nagy 2010; Wong *et al.* 2012). However, it is possible to recover high molecular weight DNA from vertebrate tissue preserved at ambient temperature for field appropriate times (Nietfeldt & Ballinger 1989; Seutin *et al.* 1991; Muralidharan & Wemmer 1994; Kilpatrick 2002; Nagy 2010; Michaud & Foran 2011). Opportunistic encounters with animal carcasses in the wild also provide sampling opportunities from which it might be possible to recover high-quality DNA.

Here, we test RNA and DNA preservation from rat tissue under different preservation conditions as if they had been collected in the field: collection of samples with appropriate field tools in the open air and mid-termed preservation (7–8 weeks and 10 months) at ambient temperature. We compared the quality and quantity of RNA extracted from various sample types preserved in a homemade nucleic acid preservation (NAP) buffer, in RNA*later* (Qiagen) or in Longmire buffer (Longmire *et al.* 1997). We also evaluated the quality and quantity of DNA extracted from samples preserved in NAP buffer, in 95% ethanol or in Longmire buffer. We tested the preservation conditions on samples commonly obtained when animals are euthanized (liver, brain and muscle) or when the animal is released (ear, tail and blood). We also studied the postmortem stability of nonpreserved DNA in samples taken from muscle left at room temperature for up to 2 weeks.

Materials and methods

Three rats (*Rattus rattus*) were captured and euthanized as pest control in a private garden (Seville province, Spain) and donated by the owners. Within 25 min after death, several samples were taken from each individual in the following order: blood from cardiac puncture, liver, brain, muscle from the hind legs, ear and tail tip. Collectors were trained to sample around 6 mm^2 from ear and 50–90 mg for the other sample types. Samples were placed in 1.5-mL Eppendorf tubes and preserved in five different ways: (i) snap-frozen in liquid nitrogen and then stored at −80 °C; (ii) 95% ethanol; (iii) Longmire buffer (Longmire *et al.* 1997); (iv) RNA*later* (Qiagen); and (v) NAP buffer. Preservation at −80 °C was used as a positive control for DNA and RNA preservation. The NAP buffer consisted of 0.019 M ethylenediaminetetraacetic acid (EDTA) disodium salt dihydrate, 0.018 M sodium citrate trisodium salt dihydrate, 3.8 M ammonium sulphate and was adjusted to pH 5.2 with H_2SO_4 (see Appendix I for the full protocol). The recipe is from the Research Coordination Network in Ecoimmunology website (www.ecoimmunology.org).

RNA preservation

The samples collected to study RNA preservation were left at ambient temperature for about 8 weeks (59–66 days) or about 10 months (294 days). We then extracted RNA from blood, liver, brain and ear samples using the PureLink™ RNA Mini Kit (Ambion, Life Technologies) and from muscle using the RNeasy Fibrous Tissue MiniKit (Qiagen), following the manufacturers' protocols. For the second time point, only liver and ear were available. They were cut in half and RNA was extracted from each in independent reactions.

RNA concentration was determined from the extracts with a NanoDrop ND-1000 Spectrophotometer (NanoDrop Technologies, Inc., Wilmington, DE, USA). The RNA quality was quantified in a 2100 BioAnalyzer (Applied Biosystems), which determines the RNA integrity number (RIN). This parameter estimates the RNA integrity on a scale from 1 (RNA is completely degraded) to 10 (RNA shows no degradation) as a function of the RNA electrophoretic profile (Schroeder *et al.* 2006).

DNA preservation

The samples collected to study DNA preservation were left at ambient temperature for about 7 weeks (49–51 days) or about 10 months (298 days). Liver, brain, muscle and tail tip samples were digested overnight at 37 °C with proteinase K (Roche Diagnostics, Germany), and then, DNA was extracted using the High Pure PCR Template Preparation Kit (Roche Diagnostics, Germany) following the manufacturer's protocol. For the second time point (10 months), only brain and tail were available. They were cut in half and the DNA of each was extracted in independent reactions. We determined DNA concentrations and quality of the extractions with a NanoDrop ND-1000 Spectrophotometer and ran 1% agarose gels stained with SYBR® SAFE (Invitrogen, USA) to assess DNA degradation.

Postmortem stability of DNA

Muscle samples were left at room temperature in 1.5-mL Eppendorf tubes and snap-frozen in liquid nitrogen at times 0 h, 2 h, 6 h, 12 h, 24 h, 48 h, 1 week and 2 weeks, after which they were stored at −80 °C. The DNA extraction, quantification and quality assessment were done as above.

Data analysis

The effects of sample type and preservation conditions on the RNA quality (RIN) and quantity and on the DNA quantity were assessed by fitting generalized linear

models, adopting Gaussian or negative binomial error distributions (with either identity or log link functions, respectively) as they best fit each variable. Tukey–Kramer post hoc tests were then conducted to test for differences among the different levels of each factor. We tested the effect of variation among rat individuals on the quantity of RNA and DNA and on the RNA quality across tissues and found it to be not significant in any case. Hence, we did not include rat identity in any further analysis. All analyses were run in SAS v. 9.1 (SAS Institute Inc., USA).

Results

RNA preservation

RNA from cryopreserved samples showed very little degradation as indicated by the BioAnalyzer. All cryopreserved samples except blood had two clear 18S and 28S peaks and high RIN values (mean ± SD: 8.6 ± 0.8; Table 1). Blood samples had a profile with two clear 18S and 28S peaks, but very low RNA concentration (mean ± SD: 20.8 ± 19.2 ng/μL; Table 1), such that the BioAnalyzer software was not able to calculate RIN values. All samples from liver, muscle, brain and ear preserved in RNA*later* and NAP buffer for 8 weeks were partially degraded as revealed by their electrophoretic profiles (i.e. Fig. 2). They all showed a clear 18S peak, but a very low or no 28S peak. We excluded two of the three muscle samples preserved in RNA*later* from our results because they showed very low electrophoretic profiles, probably due to RNA extractions that did not work properly (see discussion). Samples preserved in RNA*later* and NAP buffer experienced similar degradation for the same time point, but there was substantial RNA degradation between 8 weeks and 10 months (Figs 1 and 2; Table 1).

Extractions from cryopreserved samples had 1.5 times higher RNA concentration than those preserved in

Table 1 Mean RNA or DNA concentration (ng/μL ± SD) for each rat sample type and preservation condition. For RNA, we also report the RNA integrity number (RIN) as a measure of quality. For −80 °C, samples were snap-frozen in liquid nitrogen and then stored at −80 °C, NAP buffer as described in text, RNA*later* is a commercial product from Qiagen, the ethanol was at 95%, and Longmire refers to the lysis buffer described in Longmire *et al*. (1997). RNA extracts from blood and Longmire buffer were very degraded and/or had low concentrations, such that the BioAnalyzer could not estimate RIN values (NA) or were not run (NR). $N = 3$ samples (three different rats) for each combination of tissue × condition at times 7 or 8 weeks and for each time point in the postmortem DNA stability, except for muscle in RNA*later* ($N = 1$). For the 10-month time point, $N = 6$ (3 rats × 2 replicates per rat) for each combination of tissue × condition, except $n = 4$ for tail in ethanol

		Sample	−80 °C	NAP buffer	RNA*later*	Ethanol	Longmire
RNA preservation	8 weeks	Blood	20.8 ± 19.2 RIN: NA/NR	4.0 ± 3.1 RIN: NR	3.5 ± 2.0 RIN: NA/NR	–	15.9 ± 11.4 RIN: NR
		Liver	471.6 ± 178.4 RIN: 8.7–9.4	145.5 ± 67.7 RIN: 6.1–6.4	123.8 ± 65.1 RIN: 5.2–6.5	–	1.4 ± 1.1 RIN: NA/NR
		Brain	102.6 ± 47.4 RIN: 7.5–8.2	43.3 ± 21.4 RIN: 4.9–5.4	53.2 ± 12.7 RIN: 5.0–6.2	–	0.9 ± 0.3 RIN: NR
		Muscle	120.2 ± 109.2 RIN: 8.9–9.7	211.6 ± 29.8 RIN: 4.2–4.6	134.2 RIN: 4.8	–	49.7 ± 4.1 RIN: NA
		Ear	32.7 ± 6.5 RIN: 7.4–9.0	30.2 ± 8.1 RIN: 3.5–5.1	23.1 ± 1.5 RIN: 4.0–4.9	–	5.9 ± 5.3 RIN: NR
	10 months	Liver	590.2 ± 249.6 RIN: 7.3–9.7	383.9 ± 180.5 RIN: 2.5–2.9	271.3 ± 81.7 RIN: 2.3–3.2	–	–
		Ear	43.8 ± 39.3 RIN: 2.2–7.8	23.3 ± 7.7 RIN: 1–2.4	38.6 ± 18.1 RIN: NA-2.5	–	–
DNA preservation	7 weeks	Liver	88.8 ± 27.5	98.2 ± 45.6	–	96.0 ± 47.2	10.4 ± 4.9
		Brain	26.4 ± 9.1	73.3 ± 5.0	–	36.4 ± 10.5	14.0 ± 2.2
		Muscle	37.3 ± 5.6*	53.2 ± 7.2	–	32.8 ± 16.4	15.3 ± 8.4
		Tail	67.7 ± 18.5	65.5 ± 17.9	–	63.1 ± 16.6	16.3 ± 5.7
	10 months	Brain	16.1 ± 2.9	45.4 ± 9.9	–	31.4 ± 9.2	25.5 ± 15.6
		Tail	14.8 ± 6.7	57.5 ± 27.0	–	35.8 ± 7.9	25.3 ± 17.6

DNA postmortem stability	Muscle left at ambient temperature							
	0 h	2 h	6 h	12 h	24 h	48 h	1 week	2 weeks
	37.3 ± 5.6*	34.3 ± 2.8	42.5 ± 18.6	39.3 ± 15.6	40.8 ± 15.0	41.1 ± 20.2	42.3 ± 29.7	9.4 ± 5.1

*Same sample.

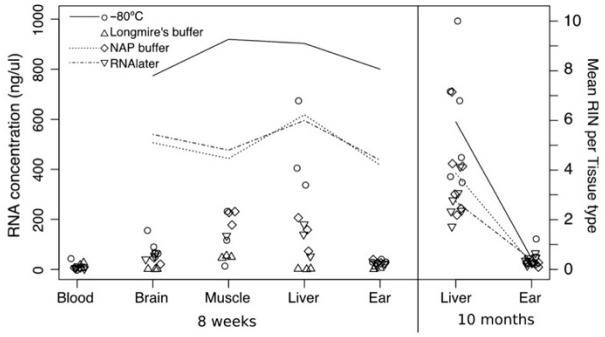


Fig. 1 RNA concentration and quality from various rat sample types under different preservation conditions. RNA concentrations (left *y*-axis) are represented with symbols. The mean of the RNA integrity number (RIN) (right *y*-axis) is represented with a line for each preservation condition. For plotting purposes, RIN values for very degraded samples were all considered 0. Two muscle replicates are not included (see text). Samples stored at −80 °C yielded the highest RNA concentrations and little degradation, reflected in high RIN values for all sample types. The samples preserved in RNA*later* and in NAP buffer were all partially degraded and had very similar quality and quantity values for any given sample type. Liver had the highest RNA concentrations, and ear the lowest. Samples in Longmire buffer and noncryopreserved blood were completely degraded.

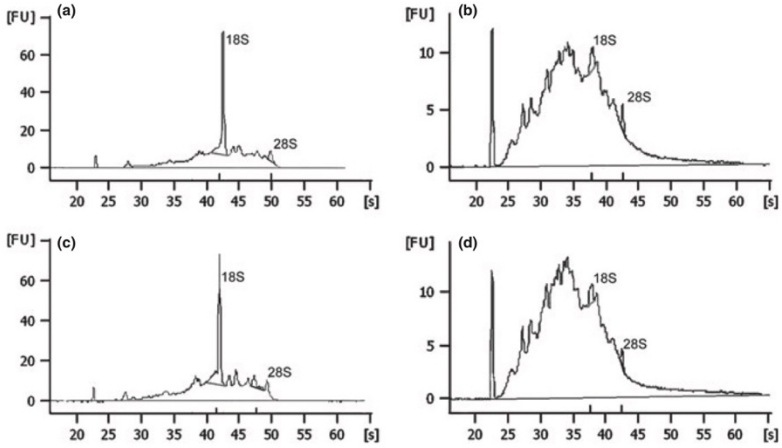


Fig. 2 BioAnalyzer profiles of RNA from liver preserved for 8 weeks or 10 months, respectively, in (a, b) RNA*later* and (c, d) NAP buffer. Samples preserved in the two different buffers had similar electrophoretic profiles within a time period, with a clear 18S peak, but a very degraded 28S peak, and similar RIN values of 6.3 and 6.1, respectively, after 8 weeks.

RNA*later* or NAP buffer for 8 weeks (Table 1; Fig. 1). Samples preserved in NAP buffer did not significantly differ in RNA concentration from those preserved in RNA*later* ($P = 0.65$). After 10 months, cryopreserved samples retained a higher RNA concentration, and RNA concentrations still did not differ between NAP buffer and RNA*later* ($P = 0.90$). After 10 months, the difference in RNA quality was much greater for cryopreserved samples than for either preserving buffer (Table 1), and so was RNA quantity.

The sample type had a significant effect on the RNA quality ($P < 0.001$) and quantity ($P < 0.001$) in samples preserved for 8 weeks. Liver was the tissue that yielded the highest RNA concentration and quality, whereas ear samples yielded the lowest (Fig. 1). Noncryopreserved blood and samples in Longmire buffer were completely degraded (Fig. 1), as indicated by their absorbance curves on NanoDrop and by their BioAnalyzer profiles (results not shown). The very low RNA concentrations registered by the NanoDrop were probably artefacts due to the concentrations being below the lower limit of sensitivity of the machine. Sample type also had a significant effect on both RNA quality ($P = 0.046$) and quantity ($P < 0.001$) after 10 months, although only liver and ear types could be compared.

DNA preservation

After both 7 weeks and 10 months at ambient temperature, all samples from all combinations of sample type and preservation methods yielded high molecular weight DNA (Fig. 3). After 7 weeks, DNA extractions from samples in 95% ethanol had degraded more than those from NAP buffer or Longmire buffer (Fig. 3). After 10 months, more degradation was observed in both brain and tail for all conditions, but high molecular weight DNA was also still present in all samples tested. Preservation condition and sample type had significant effects on DNA concentration ($P < 0.001$ in both cases). Within the 7-week samples, DNA concentration from samples preserved in Longmire buffer was 4.4 times lower than

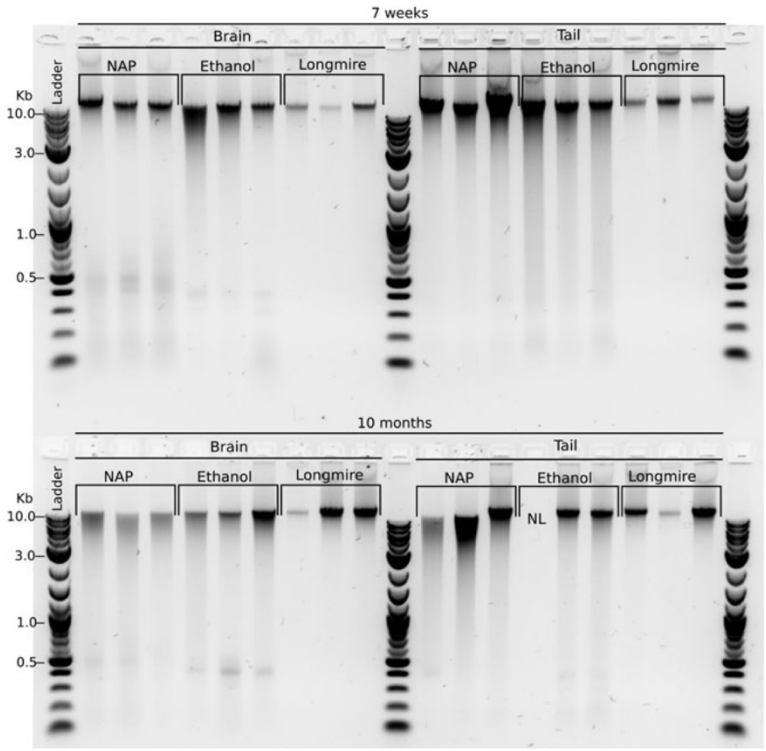


Fig. 3 DNA from brain and tail tissue preserved in NAP buffer, 95% ethanol and Longmire buffer. After 7 weeks (top row) and 10 months (bottom row) at ambient temperature, it was possible to recover high molecular weight DNA in all cases, although the samples preserved in the NAP buffer and 95% ethanol yielded brighter bands than those preserved in the Longmire buffer. Extractions from samples preserved in 95% ethanol showed more degradation than those from the NAP and Longmire buffers. A well that was not loaded is labelled 'NL'.

in the other conditions (Table 1). DNA concentrations from samples preserved in NAP buffer were 1.3 times higher than those from 95% ethanol ($P = 0.01$) and cryopreservation ($P = 0.03$) (Table 1). Liver yielded 2.1 times higher DNA concentrations than muscle ($P < 0.001$), and 1.4 times higher concentration than the tail, although this difference was not significant (Table 1). After 10 months (Table 1), DNA concentration was still significantly higher for samples preserved in NAP buffer than for those preserved in either 95% ethanol ($P = 0.043$) or cryopreserved ($P < 0.001$). We found no significant differences in DNA concentration between the two sample types compared after 10 months (tail and brain).

Postmortem stability of DNA

Agarose gels indicated that large DNA molecules persisted at high concentrations in nonpreserved muscle samples left at ambient temperature for up to 1 week (Table 1; Fig. 4). We detected DNA degradation just 6 h postmortem. After 2 weeks, none of the samples showed any high molecular weight DNA (Fig. 4).

Discussion

RNA preservation

Here, we tested several methods for RNA preservation that could be used in field expeditions where samples need to be stored at ambient temperature. RNA from animals in the field can be used for multiple types of studies, including expression analyses and transcriptome sequencing. Transcriptome sequencing is likely to be more robust to some amount of RNA degradation. Preservation of RNA under field conditions has only recently become an issue as these types of studies have become more accessible through new NGS technologies. There are relatively few publications directly addressing RNA preservation under field conditions (Table 2). Most of the studies we could find either used a commercial RNA preservation product within the specifications of that product, and subjected their samples to room temperature for only 12–72 h. There are three notable exceptions, one in which whole butterflies were preserved in RNA*later* for 10 days (Gayral *et al*. 2011), a second in which rat liver was kept in RNA*later* for 15 days (Kasahara *et al*. 2006) and another in which hairs were stored in RNA*later* for up to 12 weeks (Bradley *et al*. 2005; Table 2). Here, we showed that the economical, homemade NAP buffer was as effective as RNA*later* for preserving RNA quality and quantity for 8 weeks and 10 months. Although the preservation was the same between RNA*later* and NAP buffer, the RNA did degrade through time, and the RIN values at 10 months were much lower than those after 8 weeks. None of the conditions we tested on blood yielded RNA in useful quality and/or quantity for expression or transcriptomic studies. However, others have shown that RNA in blood is stable in several commercial products at room temperature on a much shorter timescale and can be used for NGS under those conditions (i.e. preservation time 24 h, Schwochow *et al*. 2012; Table 2) and in expression studies (i.e. preservation time 5 days, Rainen *et al*. 2002; Table 2).

DNA preservation

Several studies have reported high molecular weight or usable DNA from a variety of tissues preserved in a variety of ways compatible with extended field work, including a variety of salt- or alcohol-based solutions

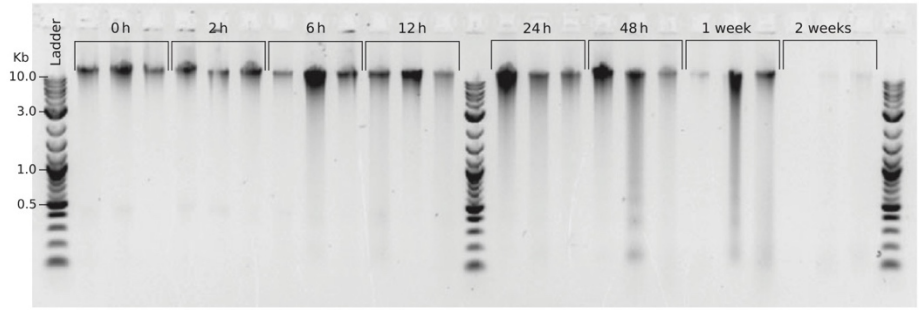


Fig. 4 DNA extractions from nonpreserved muscle tissue left at ambient temperature for up to 2 weeks. High molecular weight DNA was present in all samples from 0 h, 2 h, 6 h, 12 h, 24 h, 48 h and 1 week, but not in samples from 2 weeks. DNA degradation started to be apparent on the gel at 6 h.

Table 2 Review of previous studies analysing preserved RNA quantity or quality. Due to the small number of studies, data from all animals (vertebrate and invertebrate) were included

Tissue	Taxa	Condition	Preservation	Time	Quality test	RNA preservation	Reference
Blood	Carnivora	12–24 h at RT then −20°C	RNAlater	12–24 h	Nanodrop; BioAnalyzer	5.0–43.9 μg/500 μL blood RIN: 4.6 ± 2.3	Schwochow *et al.* (2012)
		12–24 h at RT then −20 °C	RNAprotect	12–24 h	Nanodrop; BioAnalyzer	0.0–6.2 μg/500 μL blood RIN: 6.9 ± 2.6	Schwochow *et al.* (2012)
		12–24 h at RT then −20 °C	TRIzol LS	12–24 h	Nanodrop; BioAnalyzer	0.2–15.1 μg/500 μL blood RIN: 6.2 ± 2.9	Schwochow *et al.* (2012)
		12–24 h at RT then −20 °C	PAXgene	12–24 h	Nanodrop; BioAnalyzer; Sequencing in 1/16th lanes of a 454 GS FLX Titanium	0.1–10.2 μg/500 μL blood RIN: 7.7 ± 1.2 Successful 454 sequencing of cDNA libraries from samples with RIN >7	Schwochow *et al.* (2012)
	Human	4, 20, 22 °C	PAXgene tubes	0, 1, 3, 5, 7 d	BioAnalyzer; Northern blot of GAPDH; IFN IEF SS and p53 gene transcripts	RNA integrity high for at least 5 d at 22 °C Bands in Northern blot identifiable for at least 7 d at 22 °C	Rainen *et al.* (2002)
Blood (Filtered)	Carnivora	12–24 h at RT then −20 °C	RNA later	0, 1, 3, 5, 7 d	Nanodrop; BioAnalyzer	0.1–3.7 μg/500 μL blood RIN: 7.6 ± 1.9	Schwochow *et al.* (2012)
Hair	Human	RT	RNAlater	1, 3, 6, 12 weeks	Amplification of β-actin (318-bp) and MITF (314-bp) segments from cDNA	Consistent amplification of MITF and β-actin from three hairs through 6 w	Bradley *et al.* (2005)
Liver	Mouse	RT	RNAlater	15 min, 1 h, 4 h, 24 h	18S, 28S bands on agarose gel	RNA stable for 24 h	Vinceˇk *et al.* (2003)
		RT	100% ethanol	15 min, 1 h, 4 h, 24 h	18S, 28S bands on agarose gel	RNA stable for 24 h	Vinceˇk *et al.* (2003)
		RT	0.9% NaCl	15 min, 1 h, 4 h, 24 h	18S, 28S bands on agarose gel	RNA stable for 24 h	Vinceˇk *et al.* (2003)
		RT	Xylene	15 min, 1 h, 4 h, 24 h	18S, 28S bands on agarose gel	RNA degrades in 1–4 h	Vinceˇk *et al.* (2003)
		RT	10% formalin	15 min, 1 h, 4 h, 24 h	18S, 28S bands on agarose gel	RNA degrades in <15 min	Vinceˇk *et al.* (2003)
	Rat	RT	RNAlater	1, 3, 8, 15 d	28S/18S ratio on agarose gel; β-Actin mRNA quantified by RT-PCR	Clear 18S and 28S bands through day 15 Approximately, 30% reduction in β-actin mRNA copies by day 15	Kasahara *et al.* (2006)
Whole animal	Artemia spp.	5 °C	RNAlater	0, 1, 2, 4 and 8 m	Quantification on fluorometer	Stable concentration for at least 4 m	Gorokhova (2005)
		19–22 °C	RNAlater	0, 1, 2, 4 and 8 m	Quantification on fluorometer	Stable concentration for at least 1 m	Gorokhova (2005)
	Butterfly	RT	RNAlater	10 d	BioAnalyzer; Illumina library preparation	Very good RNA quality; Successful preparation of Illumina cDNA libraries	Gayral *et al.* (2011)

min, minutes; h, hours; d, days; w, weeks; m, months; RT, room temperature.

Table 3 Survey of some studies reporting on DNA preservation in birds and mammals over weeks or months timescales

Tissue	Taxon	Condition	Time	Quality test	Results	Reference
Blood	Bird	Dried on glass	6 w	Agarose gel; Southern blot	HMW DNA	Seutin *et al.* (1991)
	Bird	Lysis buffer (Applied Biosystems)	6 w	Agarose gel; Southern blot	HMW DNA	Seutin *et al.* (1991)
	Bird	Queen's lysis buffer	24 w	Agarose gel; Southern blot	HMW DNA	Seutin *et al.* (1991)
	Human	Dried on filter paper	4.5 m	Agarose gel	HMW DNA	McCabe *et al.* (1987)
	Human	Dried on cloth	4 y	Agarose gel; Southern blot of HinfI digestion	HMW DNA; unique fingerprinting	Gill *et al.* (1985)
Peripheral blood leucocytes	Elephant	LST buffer	1, 4, 6, 8 w	PCR of mitochondrial (520-bp) and nuclear (260-bp) regions	Successful PCR's after 6 w	Muralidharan & Wemmer (1994)
Brain	Bird	DMSO	6 w	Agarose gel; Southern blot	HMW DNA	Seutin *et al.* (1991)
	Bird	Ethanol 70%	6 w	Agarose gel; Southern blot	No DNA recovered	Seutin *et al.* (1991)
Liver	Mouse	DMSO	1, 3, 5 d; 1, 2, 3, 4, 6 w; 2, 3, 4, 5, 6 m; 2 y	Agarose gel, PCR of cyt b	HMW DNA; Successful PCR	Kilpatrick (2002)
	Mouse	95% ethanol	1, 3, 5 d; 1, 2, 3, 4, 6 w; 2, 3, 4, 5, 6 m; 2 y	Agarose gel, PCR of cyt b	HMW DNA; Successful PCR	Kilpatrick (2002)
	Mouse	Longmire buffer	1, 3, 5 d; 1, 2, 3, 4, 6 w; 2, 3, 4, 5, 6 m; 2 y	Agarose gel, PCR of cyt b	HMW DNA; Successful PCR	Kilpatrick (2002)
	Bird	DMSO	6, 24 w	Agarose gel; Southern blot	HMW DNA	Seutin *et al.* (1991)
	Bird	Ethanol, 70%	6, 11 w	Agarose gel; Southern blot	Significant DNA degradation	Seutin *et al.* (1991)
Muscle	Bird	DMSO	6 w	Agarose gel; Southern blot	HMW DNA	Seutin *et al.* (1991)
	Bird	Ethanol 70%	6 w	Agarose gel; Southern blot	No DNA recovered	Seutin *et al.* (1991)
	Human	Dehydration	4, 7, 14, 28 d	STR genotyping	Full profile	Allen-Hall & McNevin (2012)
	Human	DMSO	4, 7, 14, 28 d	STR genotyping	Full profile	Allen-Hall & McNevin 2012;
	Human	DNAgard	4, 7, 14, 28 d	STR genotyping	Full profile	Allen-Hall & McNevin (2012)
	Human	Ethanol 70%	4, 7, 14, 28 d	STR genotyping	Full profile	Allen-Hall & McNevin (2012)
	Human	Ethanol 70% + 0.1 mM EDTA	4, 7, 14, 28 d	STR genotyping	Full profile	Allen-Hall & McNevin (2012)
	Human	Genotek Tissue Stabilising Kit	4, 7, 14, 28 d	STR genotyping	Full profile	Allen-Hall & McNevin (2012)
	Human	Solid NaCl	4, 7, 14, 28 d	STR genotyping	Full profile up to 7d	Allen-Hall & McNevin (2012)
	Human	RNAlater	4, 7, 14, 28 d	STR genotyping	Frequent allelic dropout	Allen-Hall & McNevin (2012)

Table 3 (Continued)

Tissue	Taxon	Condition	Time	Quality test	Results	Reference
	Human	TENT buffer	4, 7, 14, 28 d	STR genotyping	Frequent allelic dropout	Allen-Hall & McNevin (2012)
Muscle and skin	Pig	Dried at 70 °C for 72 h	2 w, 2 m	PCR of IGF-1 (642 bp)	100% PCR success	Michaud & Foran (2011)
	Pig	Ethanol 70%	2 w, 2 m	PCR of IGF-1 (642 bp)	100% PCR success	Michaud & Foran (2011)
	Pig	Isopropanol, 70%	2 w, 2 m	PCR of IGF-1 (642 bp)	100% PCR success	Michaud & Foran (2011)
	Pig	RNAlater	2 w, 2 m	PCR of IGF-1 (642 bp)	100% PCR success	Michaud & Foran (2011)
	Pig	Silica desiccant	2 w, 2 m	PCR of IGF-1 (642 bp)	100% PCR success	Michaud & Foran (2011)
	Pig	−80 °C	1 w, 2 m, 6 m	PCR of IGF-1 (642 bp)	75%; 100%; 50% PCR success	Michaud & Foran (2011)
	Pig	DMSO	1 w, 2 m, 6 m	PCR of IGF-1 (642 bp)	100%; 100%; 75% PCR success	Michaud & Foran (2011)
	Pig	Ethanol, 40%	1 w, 2 m, 6 m	PCR of IGF-1 (642 bp)	100%; 100%; 0% PCR success	Michaud & Foran (2011)
	Pig	Ethanol, 70%	1 w, 2 m, 6 m	PCR of IGF-1 (642 bp)	100%; 50%; 50% PCR success	Michaud & Foran (2011)
	Pig	Ethanol, 100%	1 w, 2 m, 6 m	PCR of IGF-1 (642 bp)	100%; 100%; 50% PCR success	Michaud & Foran (2011)
	Pig	Isopropanol, 70%	1 w, 2 m, 6 m	PCR of IGF-1 (642 bp)	100%; 100%; 25% PCR success	Michaud & Foran (2011)
	Pig	Isopropanol, 100%	1 w, 2 m, 6 m	PCR of IGF-1 (642 bp)	100%; 100%; 25% PCR success	Michaud & Foran (2011)
	Pig	Silica desiccant, 2.5 g	1 w, 2 m, 6 m	PCR of IGF-1 (642 bp)	100%; 100%; 0% PCR success	Michaud & Foran (2011)
	Pig	Silica desiccant, 12.5 g	1 w, 2 m, 6 m	PCR of IGF-1 (642 bp)	100%; 50%; 0% PCR success	Michaud & Foran (2011)

d, days; w, weeks; m, months; y, years; HMW, high molecular weight, bp, base pairs; PCR, polymerase chain reaction.

and desiccation procedures (Table 3). The NAP buffer preserved DNA quality and quantity slightly better than both 95% ethanol and cryopreservation for at least 7 weeks. This could be because EDTA, which is present in the NAP buffer, might protect the DNA during the extraction procedure. In another study, Kilpatrick (2002) showed that addition of EDTA to ethanol prevented DNA degradation during the extraction process. He also recorded that the use of salt-based buffers that contained EDTA, such as DMSO or Longmire, could preserve high molecular weight DNA after noncryogenic storage of tissue samples for long times (at least 2 years) at room temperature. Therefore, it is likely that the DNA could be stable in NAP buffer at room temperature for much longer than the 10 months demonstrated here, perhaps even years. Longmire buffer is a lysis buffer in which DNA can accumulate in the solution with time (Kilpatrick 2002). This could explain the low DNA quantity yields we obtained from tissue extractions preserved in this buffer. Nevertheless, DNA quality was high in samples preserved in Longmire buffer.

Useful quantities and qualities of DNA were observed here in nonpreserved postmortem muscle tissue for up to 1 week. Other authors have also found that high molecular weight DNA in some tissues, such as in blood or kidney, degrade very fast after 1 week, whereas in others, such as brain, lasts longer (Ludes *et al.* 1993). Although these times will vary depending on external humidity and temperature, sampling from recently dead carcasses in the field can be a potential source for high molecular weight DNA.

Sample types

RNA and DNA stability can be tissue dependent (RNA: Bahar *et al.* 2007; Seear & Sweeney 2008; DNA: Bär *et al.* 1988; Ludes *et al.* 1993). Our results show that liver yielded the best quality and quantity DNA and RNA among the sample types tested. In vertebrates, liver is the next best tissue after testis that yields the highest quantity of high molecular weight DNA (Wong *et al.* 2012). Liver also offers a lot of tissue quantity for DNA extraction, as it is a big organ, but it has the risk of nucleic acid degradation due to its high nuclease content (Wong *et al.* 2012). Skeletal muscle is less prone to DNA degradation and it is also abundant, but usually yields less DNA due to the tough nature of the muscle fibres (Wong *et al.* 2012). In this study, we observed unex-

pected electrophoretic profiles for two of our extracts of RNA from muscle, perhaps for this reason.

High-throughput sequencing platforms usually require initial input of high quantities of good-quality RNA or DNA. For library preparation, sequencing services generally request more than 1 μg of nondegraded DNA and at least 1 μg of RNA with RIN >7. Other applications, such as whole-genome sequencing, recommend even larger amounts of DNA such as 1 mg of high-quality DNA (Wong *et al.* 2012). RIN values for liver preserved for 8 weeks in RNA*later* and in NAP buffer were 5.2–6.5, slightly lower than normally recommended for transcriptome sequencing. Ear clips in RNA*later* and NAP buffer yielded RNA in low quantities (mean ± SD: 26.7 ± 6.5 ng/μL), but the quality was moderate (RIN: 3.5–5.1). As NGS technologies develop, their demand for high-quantity and high-quality material may be relaxed. For example, the new Smart-Seq can perform transcriptomic analysis from RNA quantities as low as 10 pg (Goetz & Trimarchi 2012).

Conclusion

Cryopreservation should be used whenever possible as it preserves high-quantity and good-quality RNA and DNA. However, field trips often occur in locations where cryopreservation is not possible. Under such conditions, we recommend the use of NAP buffer because it is inexpensive, easy to transport because it is nonhazardous and nonflammable, and it is possible to recover a high quantity of high molecular weight DNA and medium-quality RNA after months at ambient temperature. The limited data currently available suggest that RNA preservation varies among tissues (Fig. 1), and possibly between taxa, so it would be safest to perform a pilot study as similar to the target study as possible to determine whether a usable amount of RNA is likely to be preserved under those particular conditions. Further, NAP buffer can be used for both RNA and DNA preservation. Liver is the best source for RNA and DNA and its preservation in NAP buffer offers the potential for it to be used in NGS applications. However, if animals are not collected, the biological material that can be sampled, such as tail tip or ear clip, offers fewer possibilities for expression studies due to the low quantity of RNA, although they remain a good source for DNA.

Acknowledgements

We thank the Research Coordination Network in Ecoimmunology collaborative network and their website supported by NSF-0947177 for publishing the recipe that is tested here. We are grateful to Christophe Lejeusne for his advices on working with RNA in the laboratory and to Anna Cornellas and Lola Asencio for laboratory assistance. Logistical support was provided by Laboratorio de Ecología Molecular, Estación Biológica de Doñana, CSIC (LEM-EBD), and by the Genomics Unit, CABIMER-CSIC. This work was supported by the Spanish Ministry of Science and Innovation grants CGL-11123 to IG-M and CGL2010-21524 to JAL. MCS is supported by the Spanish Ministry of Science and Innovation Predoctoral Fellowship BES-2011-049186, and PBG is supported by the Spanish Ministry of Education, Culture and Sports Predoctoral Fellowship AP2010-5373.

References

Allen-Hall A, McNevin D (2012) Human tissue preservation for disaster victim identification (DVI) in tropical climates. *Forensic Science International: Genetics*, **6**, 653–657.

Bahar B, Monahan FJ, Moloney AP *et al.* (2007) Long-term stability of RNA in post-mortem bovine skeletal muscle, liver and subcutaneous adipose tissues. *BMC Molecular Biology*, **8**, 108.

Bär W, Kratzer A, Mächler M, Schmid W (1988) Postmortem stability of DNA. *Forensic Science International*, **39**, 59–70.

Bradley BJ, Pastorini J, Mundy NI (2005) Successful retrieval of mRNA from hair follicles stored at room temperature: implications for studying gene expression in wild mammals. *Molecular Ecology Notes*, **5**, 961–964.

Chen S, Zhou R, Huang Y *et al.* (2011) Transcriptome sequencing of a highly salt tolerant mangrove species *Sonneratia alba* using Illumina platform. *Marine Genomics*, **4**, 129–136.

Dassanayake M, Haas JS, Bohnert HJ, Cheeseman JM (2009) Shedding light on an extremophile lifestyle through transcriptomics. *The New Phytologist*, **183**, 764–775.

Elmer KR, Fan S, Gunter HM *et al.* (2010) Rapid evolution and selection inferred from the transcriptomes of sympatric crater lake cichlid fishes. *Molecular Ecology*, **19**(Suppl.1), 197–211.

Gayral P, Weinert L, Chiari Y *et al.* (2011) Next-generation sequencing of transcriptomes: a guide to RNA isolation in nonmodel animals. *Molecular Ecology Resources*, **11**, 650–661.

Gill P, Jeffreys AJ, Werrett DJ (1985) Forensic applications of DNA "fingerprints". *Nature*, **318**, 577–579.

Goetz JJ, Trimarchi JM (2012) Transcriptome sequencing of single cells with Smart-Seq. *Nature Biotechnology*, **30**, 763–765.

Gorokhova E (2005) Effects of preservation and storage of microcrustaceans in RNAlater on RNA and DNA degradation. *Limnol Oceanogr Methods*, **3**, 143–148.

Hohenlohe PA, Bassham S, Etter PD, *et al.* (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics*, **6**, e1000862.

Kasahara T, Miyazaki T, Nitta H (2006) Evaluation of methods for duration of RNA quality in rat liver used for transcriptome analysis. *The Journal of Toxicological Sciences*, **31**, 509–519.

Kilpatrick CW (2002) Noncryogenic preservation of mammalian tissues for DNA extraction: an assessment of storage methods. *Biochemical Genetics*, **40**, 53–62.

Longmire JL, Maltbie M, Baker RJ (1997) Use of "lysis buffer" in DNA isolation and its implications for museum collections. *Museum of Texas Tech University*, **163**, 1–3.

Ludes B, Pfitzinger H, Mangin P (1993) DNA fingerprinting from tissues after variable postmortem periods. *Journal of Forensic Sciences*, **38**, 686–690.

Massie H, Samis H, Baird M (1972) The kinetics of degradation of DNA and RNA by H_2O_2. *Biochimica et Biophysica Acta*, **272**, 539–548.

McCabe ER, Huang SZ, Seltzer WK, Law ML (1987) DNA microextraction from dried blood spots on filter paper blotters: potential applications to newborn screening. *Human Genetics*, **75**, 213–216.

Michaud CL, Foran DR (2011) Simplified field preservation of tissues for subsequent DNA analyses. *Journal of Forensic Sciences*, **56**, 846–852.

Muralidharan K, Wemmer C (1994) Transporting and Storing field-collected specimens for DNA without refrigeration for subsequent DNA extraction and analysis. *BioTechniques*, **17**, 420–422.
Nagy ZT (2010) A hands-on overview of tissue preservation methods for molecular genetic analyses. *Organisms Diversity & Evolution*, **10**, 91–105.
Nietfeldt J, Ballinger R (1989) A new method for storing animal tissue prior to mtDNA extraction. *BioTechniques*, **7**, 31–32.
Rainen L, Oelmueller U, Jurgensen S *et al.* (2002) Stabilization of mRNA expression in whole blood samples. *Clinical Chemistry*, **48**, 1883–1890.
Riesgo A, Pérez-Porro AR, Carmona S, Leys SP, Giribet G (2012) Optimization of preservation and storage time of sponge tissues to obtain quality mRNA for next-generation sequencing. *Molecular Ecology Resources*, **12**, 312–322.
Schroeder A, Mueller O, Stocker S *et al.* (2006) The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Molecular Biology*, **7**, 3.
Schwochow D, Serieys LEK, Wayne RK, Thalmann O (2012) Efficient recovery of whole blood RNA - a comparison of commercial RNA extraction protocols for high-throughput applications in wildlife species. *BMC Biotechnology*, **12**, 33.
Seear PJ, Sweeney GE (2008) Stability of RNA isolated from post-mortem tissues of Atlantic salmon (Salmo salar L.). *Fish Physiology and Biochemistry*, **34**, 19–24.
Seutin G, White B, Boag P (1991) Preservation of avian blood and tissue samples for DNA analyses. *Canadian Journal of Zoology*, **69**, 89–92.
Vincek V, Nassiri M, Knowles J (2003) Preservation of Tissue RNA in Normal Saline. *Laboratory Investigation*, **83**, 137–138.
Wang S, Sherman M (2006) Cervical tissue collection methods for RNA preservation: comparison of snap-frozen, ethanol-fixed, and RNAlater-fixation. *Diagnostic Molecular Pathology*, **15**, 144–148.
Wolf JBW, Lindell J, Backström N (2010) Speciation genetics: current status and evolving approaches. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **365**, 1717–1733.
Wong PB, Wiley EO, Johnson WE *et al.* (2012) Tissue sampling methods and standards for vertebrate genomics. *GigaScience*, **1**, 8.

J.A.L., P.B. and M.C. conceived the study. I.G.M., J.A.L., P.B. and M.C. designed the experiment. P.B. and M.C. did laboratory work. I.G.M. did statistical analyses. M.C. wrote the manuscript with help from J.A.L., and P.B. and I.G.M. revised it.

Data Accessibility

Concentration of DNA and RNA, quality of RNA (RIN values) and concentration of DNA extracted across post-mortem time series are available on DRYAD repository (doi:10.5061/dryad.8gh7p).

Appendix I

Protocol for the preparation of Nucleic Acid Preservation (NAP) Buffer

Materials	Equipment
EDTA disodium salt dihydrate	Scale
Sodium citrate trisodium salt dihydrate	Weigh boat or paper
Ammonium sulfate	Magnetic stirrer with heating plate
Ultra-purified, molecular grade water	Stirring rod
H_2SO_4 to adjust the pH	PH reader
bottle or flask	

To make NAP buffer:

1. Combine 7.44 g of EDTA, 7.35 g of sodium citrate trisodium salt dihydrate, and 700 g of ammonium sulfate in 1 L of water in bottle or flask. Stir on low to moderate heat until the ammonium sulfate dissolves completely, which usually takes hours.
2. Cool to room temperature, then adjust pH to 5.2 with H_2SO_4.
3. Store at room temperature or keep refrigerated until aliquoted.
4. Aliquot 1.5 mL of buffer into 2 mL tubes for preservation of up to 150 mg of sliced tissue.

Appendix 2: PUBLICATION. Effect of the enzyme and PCR conditions of the quality of high-throughput DNA sequencing results

SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
PCR-BASED TECHNIQUES
GENOTYPING AND HAPLOTYPING
GENETIC VARIATION

Received
22 August 2014

Accepted
2 January 2015

Published
27 January 2015

Correspondence and requests for materials should be addressed to J.A.L. (jleonard@ebd.csic.es)

* These authors contributed equally to this work.

Effect of the enzyme and PCR conditions on the quality of high-throughput DNA sequencing results

Claudia Brandariz-Fontes[1,2]*, Miguel Camacho-Sanchez[1]*, Carles Vilà[1], José Luis Vega-Pla[3], Ciro Rico[4] & Jennifer A. Leonard[1]

[1]Conservation and Evolutionary Genetics Group, Estación Biológica de Doñana (EBD-CSIC), Sevilla, Spain, [2]Facultad de Medicina Veterinaria, Universidad de Panamá, Panama, [3]Laboratorio de Investigación Aplicada, Cría Caballar de las Fuerzas Armadas, Córdoba, Spain, [4]University of the South Pacific, School of Marine Studies, Faculty of Science, Technology & Environment, Laucala Campus, Suva, Fiji.

Library preparation protocols for high-throughput DNA sequencing (HTS) include amplification steps in which errors can build up. In order to have confidence in the sequencing data, it is important to understand the effects of different *Taq* polymerases and PCR amplification protocols on the DNA molecules sequenced. We compared thirteen enzymes in three different marker systems: simple, single copy nuclear gene and complex multi-gene family. We also tested a modified PCR protocol, which has been suggested to reduce errors associated with amplification steps. We find that enzyme choice has a large impact on the proportion of correct sequences recovered. The most complex marker systems yielded fewer correct reads, and the proportion of correct reads was greatly affected by the enzyme used. Modified cycling conditions did reduce the number of incorrect sequences obtained in some cases, but enzyme had a much greater impact on the number of correct reads. Thus, the coverage required for the safe identification of genotypes using one of the low quality enzymes could be seven times larger than with more efficient enzymes in a biallelic system with equal amplification of the two alleles. Consequently, enzyme selection for downstream HTS has important consequences, especially in complex genetic systems.

High throughput DNA sequencing (HTS) has dramatically reduced the cost per base sequenced[1]. HTS technologies, however, are fundamentally different from Sanger sequencing and face different problems. In HTS single molecules of DNA yield sequences, as opposed to a large pool of molecules in Sanger sequencing. This exposes errors that can occur during library preparation. For example, errors could result from the misincorporation of nucleotides during the amplification steps of library preparation. During amplification there can be partial synthesis of a DNA strand that can act as a primer in a downstream polymerase chain reaction (PCR) cycle and form a chimeric sequence if it amplifies a related allele. These sources of errors originating in PCR amplification are poorly characterized, but increasingly recognized as a problem[2,3].

Recent technical advances in HTS yielding longer reads of 350 to 1000 base pairs (bp) and methodological advances such as the incorporation of index sequences allow multiple targeted loci from many individuals to be sequenced simultaneously[4–6]. Targeted loci could have different characteristics. The simplest systems, such as loci in the mitochondrial DNA, Y chromosome (in mammals) or W chromosome (in birds) loci, are expected to yield a single haplotype and are thus the easiest to determine the sequence of. Most single copy nuclear markers, which are potentially biallelic in diploid organisms, are more challenging to accurately genotype. Very complex systems, such as gene families in which many different alleles could be present in a single individual, can be very difficult to accurately characterize. PCR based errors have been shown to be a problem in the characterization of polygenic immune system loci in model organisms[2,3]. Accurately genotyping complex loci in non-model systems for which there is not a lot of comparative data to verify results can be even more challenging[7–10].

One factor that could play an important role in identifying correct alleles and genotypes using HTS approaches is the enzyme used in the DNA amplification. In this study we tested the ability of thirteen different enzymes to yield the true sequence(s) via HTS in three genetic marker systems of different complexity. We also tested if modified PCR conditions could increase the yield of correct templates, as suggested in previous studies[11–15]. Understanding the frequency and potential sources of erroneous sequences is of prime importance for the design

of optimal protocols in HTS approaches to characterize genetic diversity in individuals and populations, and is even more critical in non-model systems.

Results

We tested the ability of 13 different enzymes to yield the true sequence(s) in three different marker sets of varying complexity (see Methods, Table 1 for abbreviations). The three sets we used were: Test 1, mitochondrial DNA from wolves, expected to yield a single sequence per individual; Test 2, MHC class II exon 2 (MHC II) in horses, a single copy nuclear gene with one or two alleles per individual; and Test 3, MHC class I exon 3 (MHC I) in horses, a multigene family which could yield several alleles per individual. Three different individuals were included in each test. A further two tests (Tests 2b and 3b) were designed to evaluate the ability of modified PCR cycling conditions to reduce amplification-associated errors. These tests were done only with the two more complex systems: MHC II for Test 2b and MHC I for Test 3b.

Error patterns and rates can vary between sequencing platforms[1,16], and even independent runs in the same platform can have an effect on the genotypes[17]. Here we focus on the performance of different polymerase on a single platform in order to more reliably assess to what degree this is an important factor to take into account when designing experiments. We chose the Roche 454 Junior sequencing platform. This platform is appropriate for this experiment because it allows relatively long and variable read lengths, so the entire length of the three different PCR products could be sequenced simultaneously in single reads.

Six enzymes (Phusion, Gold, FastStart, Roche Taq, HotStar and Biotaq) worked across all tests, five of them (Velocity, OneTaq, Imax, KapHF and Pwo) worked inconsistently in different tests. We were not able to get Vent or DeepVent to amplify in any of the systems after 12–29 tries each for Tests 1–3. The sequencing run produced 102,484 reads, from which 63,942 passed size (full length) and quality filters (complete MID and primer sequences) and could be successfully assigned to the experimental units (Genetic system/Enzyme/PCR condition/Biological replica) yielding an average coverage of 566, although with a large variation (standard deviation, s. d. = 1900). The average coverage for the sequences used in Test 1 was 1004 (s. d. = 3225), 203 in Test 2 (s. d. = 194), 370 in Test 3 (s. d. = 484), 834 in Test 2b (s. d. = 2300) and 337 in Test 3b (s. d. = 546). Eleven of the 13 enzymes tested yielded a band of the expected size in Test 1, eight in Test 2, eight in Test 3, six in Test 2b and six in Test 3b.

There was a significant effect of the enzyme on the quality of the sequences obtained (proportion of reads with a correct sequence) for all tests ($p < 0.001$ in all cases). In general, Biotaq produced the lowest portion of correct reads across all tests whereas Phusion, Pwo and KapHF worked best (Supplementary File 1). For Test 1 (with only one allele expected per individual), all the enzymes that successfully amplified DNA (11 out of 13) yielded from 50–53% (OneTaq and Biotaq) to 88–92% (Phusion, Pwo and KapHF) correct reads (Figure 1). For Test 2, the proportion of correct reads was on average 23% lower than for Test 1. There was also more variation between the enzymes, with correct reads ranging from 2% (Biotaq) to 84% (Phusion) (Figure 1). For Test 3, the multigene family marker system, the recovery of correct sequences ranged from 17–20% (Biotaq, HotStar and Roche *Taq*) to 65–71% (Phusion and FastStart) (Figure 1).

For the system with up to 2 alleles, the modified PCR had no effect on the proportion of correct reads ($p = 0.31$, Test 2 vs Test 2b). For the complex system, the multigene family, the proportion of correct reads was significantly higher under the modified PCR conditions, by an average of 7.5% ($p < 0.001$, Test 3 vs Test 3b).

We used the proportion of correct sequences obtained with each enzyme from Tests 1, 2 and 2b to calculate the probability of obtaining three or more copies of the correct allele(s). We simulated this for a simple system, a haplotype (data from Test 1), and for a more complex system, a single locus with two alleles (combined data from Tests 2 & 2b). Unequal amplification of alleles in PCR reactions where more than one allele are amplified has been observed widely[18,19]. For this reason we also simulated the number of reads needed to reach the same level of confidence when one allele in the two allele system amplified twice as well as the other. For the haplotype, between 7 (for Phusion, Pwo and KapHF) and 16 (Biotaq) reads were enough to have a 99.9% probability of obtaining 3 or more correct sequences (Table 1, Figure 2A). However, the number of reads required increased sharply as the gene system got more complex. For two alleles that amplify equally, between 42 (for Phusion) and 271 (Biotaq) reads were needed to have 99.9% confidence of getting three correct copies for each of the two alleles (Table 1, Figure 2B). In the case of unequal amplification, the coverage necessary increased to 87 for Phusion, and to 395 for Biotaq (Table 1, Figure 2C).

Discussion

The *Taq* polymerase enzyme used in the PCR steps of library preparation for HTS had a very important impact on the proportion of correct reads after sequencing. In the simplest case of a single allele being present, as in mitochondrial DNA or sex specific chromosome markers (Test 1), the majority of the reads (50–92%, depending on

Table 1 | Coverage necessary to reach a 99.9% probability of recovering three copies of the correct sequence for all alleles (based on the proportion of correct reads). Since not all alleles in a PCR amplify equally well, we calculated the coverage needed when two alleles amplify at the same rate (equal amplification), and when one allele yields twice as many products as the other (unequal amplification). Enzymes that did not amplify are marked n.a. and those which amplified but for which there was insufficient data to calculate coverage are labeled i.d. Abbreviations are those used in Figures 1 and 2 and the text

Enzyme	Abbreviation	Test 1	Test 2 equal amplification	Test 2 unequal amplification
Phusion® High Fidelity DNA Polymerase (Finnzymes)	Phusion	7	42	87
KAPA HiFi™ (Kapa Biosystems)	KapHF	7	i.d.	i.d.
Pwo® DNA Polymerase (Roche)	Pwo	7	n.a.	n.a.
AmpliTaq Gold® (Applied Biosystems)	Gold	9	48	88
i-MaxTM II DNA Polymerase (iNtRON Biotechnology)	iMax	11	57	99
Taq DNA Polymerase (Roche)	Roche Taq	11	120	185
Velocity DNA Polymerase (Bioline)	Velocity	12	n.a.	n.a.
HotStarTaq® DNA Polymerase (Qiagen)	HotStar	14	97	152
FastStart® High Fidelity PCR System (Roche)	FastStart	14	45	86
Biotaq® (Bioline)	Biotaq	16	271	395
OneTaq™ DNA Polymerase (New England Biolabs)	OneTaq	i.d.	n.a.	n.a.
Vent® DNA Polymerase (New England Biolabs)	Vent	n.a.	n.a.	n.a.
Deep Vent® DNA Polymerase (New England Biolabs)	DeepVent	n.a.	n.a.	n.a.

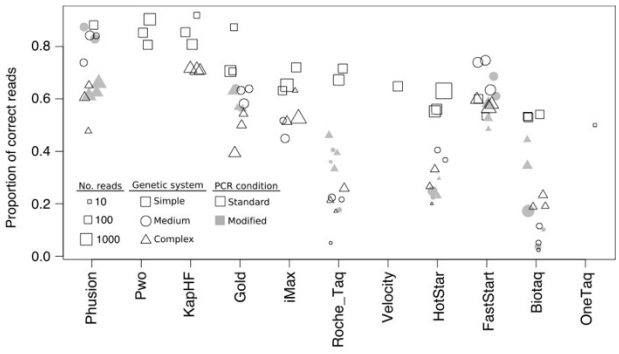


Figure 1 | **Proportion of correct reads for the three genetic systems (simple: a single allele per individual, *squares*; medium: two alleles, *circles*; and complex: multiple alleles, *triangles*) using standard PCR conditions (*open*) and modified PCR conditions to reduce chimera formation (*gray*).** The size of the shape is indicative of the number of reads (see legend). All enzymes yielded at least 50% correct reads in the simplest system, mitochondrial DNA (Test 1; open squares). Some enzymes only worked for a given set of conditions (cycling conditions/genetic system). A group of enzymes consisting of Phusion, Gold and FastStart yielded a high proportion of correct reads cosistently accross all conditions. Others, such as Roche Taq, HotStar and Biotaq, yielded a low percent of correct reads for the more complex systems (MHC class I and MHC class II). Abbreviations as defined in Table 1.

the enzyme) for all enzymes that worked (11 out of 13) had the correct sequence. In this marker system, high confidence that the haplotype identified is accurate was achieved with a low coverage of 7× for the best enzymes and 16× for the worst.

However, the proportion of correct reads went down in multi-allelic systems. In the just slightly more complex system of a single copy nuclear gene with two alleles (Test 2), the proportion of correct reads went down by an average of 23% (Figure 1). Calculations based on the proportion of sequences with the correct sequence revealed that for the best enzymes, and assuming equal amplification of the two alleles, 42 to 48 reads are necessary to have a high confidence in the identification of genotypes (probabilities of 99.9% or higher for the identification of each allele). The coverage required for the worse enzymes was much larger (above 270×).

This difference became even more pronounced when the model was more realistic and one allele amplified twice as well as the other. In this case, the coverage necessary to reach a similar degree of confidence in the results as for equal amplification of the alleles almost doubled. Differences in the amplification success of the two alleles in a biallelic system can realistically be much larger than a ratio of 1 : 2. Thus, the coverage necessary to have high confidence in a genotype would also be much higher. Cycling conditions also had a significant effect on the proportion of correct reads in some cases, but the effect was of a much smaller magnitude than the effect of the enzymes.

The results presented here suggest that for the best enzymes, under the most favorable PCR amplification conditions, and perfectly equal amplification of the two alleles of a single copy nuclear gene in a diploid organism, 42 to 48 reads are necessary to have a high confidence in the identification of genotypes. For other enzymes and in the more realistic case of unequal amplification of the alleles, nearly 400 reads are necessary to reach a similar degree of confidence in the results. This greatly complicates the analysis of data because as the number of incorrect reads goes up, the probability of these incorrect reads also being present in multiple copies also goes up. In the case of a single copy nuclear gene in a diploid organism, the maximum number of alleles that could be present is two, which reduces the bioinformatic problem. In the case where there are an unknown number of alleles, such as for MHC I, it may not be possible to determine the real alleles even with very high coverage because the frequency of reads that represent errors may grade into the frequency of reads reflecting real alleles with poor relative amplification success.

Illumina and Ion Torrent platforms are increasing their read lengths, and are now or soon will be useful for sequencing through entire PCR products. Each platform has a different rate and pattern of errors[1,20,21]. The sequences analyzed here were generated on the Roche 454 platform. However, we expect the observed large differences in enzyme performance to be evident on the other HTS platforms as well, although the exact coverage required for high confidence in the results will likely be different. New library preparation protocols that target loci without being based on PCR, such as hybridization-based enrichment, are being developed. However, they still require PCR enrichment steps, so even with these protocols, enzyme choice is important and can affect sequencing results.

Ideally, the necessary coverage for a particular system should be calculated based on the observed bias in allele amplification and errors in the enzyme and platform combination used in a particular experiment. In planning a HTS project it is also important to keep in mind that the numbers for coverage presented here to have confidence in a particular haplotype or genotype are not average numbers for a project, but minimum coverage numbers for each sample in a study. Since the coverage of all individuals analyzed simultaneously in a run is never exactly equal, the average coverage that should be planned for in a study would thus be higher.

Methods

Samples. We used DNA samples from three gray wolves (*Canis lupus*) from which the 5′ end of the mitochondrial control region had been Sanger sequenced in previous studies, and thus was known[22,23]. The loci had different GC content, from 44% to 66%, and the longest homopolymer was 5 bp (present in at least one allele of each locus). Three Retuertas breed domestic horses (*Equus caballus*) with known MHC genotypes (Brandariz-Fontes *et al.* in preparation) were selected for the nuclear loci tests. Each DNA sample was quantified using a NanoDrop ND-1000 Spectrophotometer (NanoDrop Technologies, Inc., Wilmington, DE, USA), and the concentration was adjusted to 10 or 30 ng/μl for subsequent PCR amplifications.

Taq polymerase. A range of 13 high fidelity, regular, economy and premium *Taq* polymerase enzymes were selected: Biotaq® (Bioline, London, UK), FastStart® High Fidelity PCR System (Roche, Mannheim, Germany), AmpliTaq Gold® (Applied Biosystems, Warrington, UK), HotStarTaq® DNA Polymerase (Qiagen, Hilden, Gernamy), Phusion® High Fidelity DNA Polymerase (Finnzymes, Espoo, Finland), *Taq* DNA Polymerase (Roche, Maylan, France), i-Max™ II DNA Polymerase (iNtRON Biotechnology, Seongnam, Korea), KAPA HiFi™ (Kapa Biosystems, Boston, USA), One*Taq*™ DNA Polymerase (New England Biolabs, Hitchin, UK),

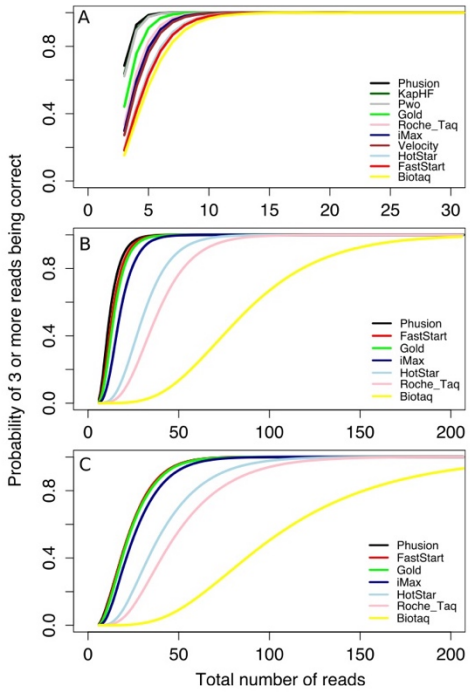


Figure 2 | Probability of obtaining 3 or more correct sequences for a given number of reads based on the proportion of correct reads observed for each enzyme and genetic system. (A). For the simplest genetic system, with only one allele per individual. (B). For a locus with two alleles that amplify equally well (3 or more correct sequences for each of the two alleles). (C). For a locus with two alleles where one amplifies twice as well as the other. Note that the scale on the X-axis in panel A is different from that in B and C.

Vent® DNA Polymerase (New England Biolabs, Hitchin, UK), Deep Vent® DNA Polymerase (New England Biolabs, Hitchin, UK), Pwo® DNA Polymerase (Roche, Maylan, France) and Velocity DNA Polymerase (Bioline, London, UK) (abbreviated names in Table 1). The list price of these enzymes for the amount recommended for a single 10 µl reaction (not including tax, handling or shipping) ranged from €0.01 to €0.63 (Spain, June 2013).

Assessment of accuracy for different enzymes. Loci for Tests 1–3 were amplified in a two-step process following the universal tailed amplicon design proposed by Roche[24,25]. First, loci were amplified with locus-specific primers with an M13 tail, and then a Multiplex Identifier (MID) and the sequencing primer were added in a second-round PCR using the same enzyme as for the first PCR. For Test 1, the 5′ end of the wolf mitochondrial control region was amplified with the primers Thr-L[22] and ddl5[26], which target a 168–172 bp fragment excluding primers (variation due to indels). For Test 2, a 257 bp fragment of MHC II in horse was amplified with primers Be3 and Be4[27]. For Test 3, a 184 bp fragment of MHC I was amplified using primers PpLAa2U270 and Ppa2L542[28]. In the second PCR, we used the first PCR as a template with a 52 bp primer which included the M13, a sample-specific 10 bp MID, the 454 Sequencing System Primer sequence and a 4 bp primer key (Table 2).

All reactions were prepared in 10 µl using the standard PCR conditions following the manufacturer's protocols that came with each enzyme for both PCR steps. These were 40 cycles of: 15 or 30 seconds at 94–98°C, 20, 30 or 90 seconds at 58°C, and 30, 60 or 90 seconds at 72°C; with a final extension at 72°C for 5, 7 or 10 minutes. All cycling was performed on a DNA Engine Peltier Thermal Cycler. All reactions, including blank controls, were checked for amplification success on a 1.5% agarose gel and visualized with SYBR®Safe (Invitrogen, Paisley, UK). All successful first PCR products were diluted and used as templates for the second-round PCRs. Second PCR products were cleaned using Agencourt AMPure xp system (Beckman Coulter, Brea, CA, USA).

Assessment of PCR protocols to reduce amplification errors. We repeated Tests 2 and 3 with modified cycling conditions in an attempt to reduce errors: Test 2b & 3b, respectively. The goal was to generate comparable data to evaluate the effect of the cycling conditions on the accuracy of the sequences (Test 2 vs 2b; Test 3 vs 3b). For the first and second PCRs the number of cycles was reduced to 25, the elongation time within cycles increased to 180 seconds and the final extension step was eliminated. Similar amplifications conditions have been suggested previously in the literature to reduce errors during amplification steps[13,14,29–32]. All the cycling reactions were performed on a DNA Engine Peltier Thermal Cycler. All reactions, including blank controls, were checked for amplification success on a 1.5% agarose gel and visualized with SYBR®Safe (Invitrogen). All successful first PCR products were diluted and used as templates for the second-round PCRs. Second PCR products were cleaned using Agencourt AMPure xp system (Beckman Coulter, Brea, CA, USA).

Library preparation and sequencing. Purified PCR products from all tests were quantified using Quant-it PicoGreen dsDNA Assay Kit (Invitrogen) in a Light Cycler 480 II real-time PCR machine (Roche). Then they were adjusted to equimolar concentration ($2x10^5$ molecules/µl in TE buffer) and all amplification products were pooled together. The pool was then quantified using the Quant-it PicoGreen dsDNA Assay Kit (Invitrogen) on a QuantiFluor™-ST fluorometer (Promega, US).

Emulsion PCR was performed according to the manufacturer's instructions with GS Junior Titanium emPCR Kit Lib-A (Roche) and sequenced in a single 454 Roche Junior run.

Data Analysis. Reads containing the complete target primers and barcodes were extracted from the multifasta output file and de-multiplexed on the basis of the barcode and loci specific primer sequences using jMHC[33]. The different sequences were compared to the known haplotype or genotype to determine correct sequences in Geneious v6.1.7 (Biomatters, Auckland, NZ). These previously known sequences were the reference against which the sequences identified in jMHC were compared, and reads were considered to have the correct sequence when it was identical to the reference. The proportion of correct reads was calculated by dividing the number of reads with correct sequences by the total number of reads from a particular amplicon.

Statistical analysis. We evaluated the effect of enzymes on the proportion of correct reads with generalized linear mixed models (GLMM), using the function lmer from the *lme4* package[34] in R (Bates, D., Maechler, M. & Bolker, B. lme4: Linear mixed-effects models using S4 classes. R package version 0.999999-0. (2012); R Core Team R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/ (2013)) for each test separately. Only cases with more than 10 reads per individual test were included in the analysis. The *Taq* polymerase was included as a fixed effect and individual as a random effect. We also used a GLMM with lmer function to evaluate the effect of the PCR protocol (standard/modified) on the proportion of correct reads for the medium and complex systems (Test 2 vs Test 2b; Test 3 vs Test 3b). PCR condition was

Table 2 | Primers used in first and second round reactions for all tests. We used published primers (references in text) upon which an M13 tail was added (shown in lower case). MIDs 1–96[25] were used in both the forward and reverse primers

Test	Primer	Sequence 5′ – 3′
Test 1	Thr-L-t	gttttcccagtcacgacGAATTCCCCGGTCTTGTAAACC
Test 1	ddl5-t	aacagctatgaccatgCATTAATGCACGACGTACATAGG
Test 2	PpLAa2U270 -t	gttttcccagtcacgacGCTTCTCATCCTAGTTCCCTT
Test 2	Ppa2L542-t	aacagctatgaccatgGCCTAGGAGTGCAGCAGA
Test 3	Be3-t	gttttcccagtcacgacGGGTCTCACACCYKCCAG
Test 3	Be4-t	aacagctatgaccatgGMGCWGCAGSGTCTCYTT
Second round	forward	CGTATCGCCTCCCTCGCGCCATCAG[*MID*]gttttcccagtcacgac
Second round	reverse	CTATGCGCCTTGCCAGCCCGCTCAG[*MID*]aacagctatgaccatg

included as a fixed effect, and enzyme and individual as random effects. We tested the significance of the variables by comparing different models using ANOVAs.

We prepared a script in Python 2.7.4 to calculate the probability of obtaining a minimum of three reads with the correct sequences for the different *Taq* enzymes when varying the total number of reads for a single haplotype and for one locus with two alleles (the script is available in Supplementary file 2, online). These probabilities are based on the frequency of correct reads observed per enzyme in Test 1 for the case of the single haplotype, and Tests 2 & 2b combined for the case of one locus with two alleles. Simulations were run only on datasets with >10 reads. We considered the number of reads when this probability reached 99.9% as an indication of the coverage needed with a given enzyme to be able to reliably identify the correct haplotype or alleles in a genotype. Often, the different alleles in multi-allelic systems do not amplify equally within a reaction. For this reason we also calculated the probability of obtaining three reads with the correct sequence for each allele when one amplifies half as well as the other.

1. Glenn, T. C. Field guide to next-generation DNA sequencers. *Mol. Ecol. Resour.* **11**, 759–769 (2011).
2. Brodin, J. *et al.* PCR-induced transitions are the major source of error in cleaned ultra-deep pyrosequencing data. *PLoS One* **8**, e70388 (2013).
3. Shugay, M. *et al.* Towards error-free profiling of immune repertoires. *Nat. Methods* **11**, 653–5 (2014).
4. Meyer, M. & Kircher, M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* **2010**, pdb.prot5448 (2010).
5. Rohland, N. & Reich, D. Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Res.* 939–946; DOI:10.1101/gr.128124.111.1 (2012).
6. Hancock-Hanser, B. L. *et al.* Targeted multiplex next-generation sequencing: advances in techniques of mitochondrial and nuclear DNA sequencing for population genomics. *Mol. Ecol. Resour.* **1**, 1–15 (2013).
7. Babik, W., Taberlet, P., Ejsmond, M. J. & Radwan, J. New generation sequencers as a tool for genotyping of highly polymorphic multilocus MHC system. *Mol. Ecol. Resour.* **9**, 713–9 (2009).
8. Babik, W. Methods for MHC genotyping in non-model vertebrates. *Mol. Ecol. Resour.* **10**, 237–51 (2010).
9. Bernatchez, L. & Landry, C. MHC studies in nonmodel vertebrates: what have we learned about natural selection in 15 years? *J. Evol. Biol.* **16**, 363–377 (2003).
10. Sommer, S. The importance of immune gene variability (MHC) in evolutionary ecology and conservation. *Front. Zool.* **2**, 16 (2005).
11. Thompson, J. R. Heteroduplexes in mixed-template amplifications: formation, consequence and elimination by "reconditioning PCR." *Nucleic Acids Res.* **30**, 2083–2088 (2002).
12. Kanagawa, T. Bias and artifacts in multitemplate polymerase chain reactions (PCR). *J. Biosci. Bioeng.* **96**, 317–323 (2003).
13. Acinas, S. G., Sarma-Rupavtarm, R., Klepac-Ceraj, V. & Polz, M. F. PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample. *Appl. Environ. Microbiol.* **71**, 8966–9 (2005).
14. Lenz, T. L. & Becker, S. Simple approach to reduce PCR artefact formation leads to reliable genotyping of MHC and other highly polymorphic loci — Implications for evolutionary analysis. *Gene* **427**, 117–123 (2008).
15. Holcomb, C. L. *et al.* Next-generation sequencing can reveal in vitro-generated PCR crossover products: Some artifactual sequences correspond to HLA alleles in the IMGT/HLA database. *Tissue Antigens* **83**, 32–40 (2014).
16. Bolotin, D. A. *et al.* Next generation sequencing for TCR repertoire profiling: Platform-specific features and correction algorithms. *Eur. J. Immunol.* **42**, 3073–3083 (2012).
17. Lighten, J., van Oosterhout, C., Paterson, I. G., Mcmullan, M. & Bentzen, P. Ultra-deep Illumina sequencing accurately identifies MHC class IIb alleles and provides evidence for copy number variation in the guppy (*Poecilia reticulata*). *Mol. Ecol. Resour.* **14**, 753–767 (2014).
18. Polz, M. F. & Cavanaugh, C. M. Bias in Template-to-Product Ratios in Multitemplate PCR Bias in Template-to-Product Ratios in Multitemplate PCR. *Appl. Environ. Microbiol.* **64**, 3724–3730 (1998).
19. Wagner, A. *et al.* Surveys of Gene Families Using Polymerase Chain Reaction: PCR Selection and PCR Drift. *Syst. Biol* **43**, 250–261 (1994).
20. Harismendy, O. *et al.* Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol.* **10**, R32 (2009).
21. Quail, M. A. *et al.* A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* **13**, 341 (2012).
22. Vila, C. *et al.* Mitochondrial DNA phylogeography and population history of the grey wolf *Canis lupus*. *Mol. Ecol.* **8**, 2089–2103 (1999).
23. Koblmüller, S., Nord, M., Wayne, R. K. & Leonard, J. A. Origin and status of the Great Lakes wolf. *Mol. Ecol.* **18**, 2313–26 (2009).
24. Daigle, D., Simen, B. B. & Pochart, P. High-throughput sequencing of PCR products tagged with universal primers using 454 life sciences systems. *Curr. Protoc. Mol. Biol.* **96**, 7.5:7.5.1–7.5.14 DOI:10.1002/0471142727.mb0705s96 (2011).
25. Roche. *Roche Technical Bulletin No. 005–2009.* (2009).
26. Leonard, J. A. *et al.* Ancient DNA evidence for Old World origin of New World dogs. *Science* **298**, 1613–6 (2002).
27. Albright-Fraser, D. G., Reid, R., Gerber, V. & Bailey, E. Polymorphism of DRA among equids. *Immunogenetics* **43**, 315–317 (1996).
28. Aldridge, B. M. *et al.* Paucity of class I MHC gene heterogeneity between individuals in the endangered Hawaiian monk seal population. *Immunogenetics* **58**, 203–15 (2006).
29. Meyerhans, A., Vartanian, J.-P. & Wain-Hobson, S. DNA recombination during PCR. *Nucleic Acids Res.* **18**, 1687–1691 (1990).
30. Judo, M. S. B., Wedel, A. B. & Wilson, C. Stimulation and suppression of PCR-mediated recombination. *Nucleic Acids Res.* **26**, 1819–1825 (1998).
31. Zylstra, P., Rothenfluh, H. S., Weiller, G. F., Blanden, R. V. & Steele, E. J. PCR amplification of murine immunoglobulin germline V genes: strategies for minimization of recombination artefacts. *Immunol. Cell Biol.* **76**, 395–405 (1998).
32. Lahr, D. J. G. & Katz, L. A. Reducing the impact of PCR mediated recombination in olecular evolution and environmental studies using a new-generation high-fidelity DNA polymerase. *Biotechniques* **47**, 857–866 (2009).
33. Stuglik, M. T., Radwan, J. & Babik, W. jMHC: software assistant for multilocus genotyping of gene families using next-generation amplicon sequencing. *Mol. Ecol. Resour.* **11**, 739–42 (2011).

Acknowledgments

The authors gratefully acknowledge Alejandro Gonzalez Voyager, Eloy Revilla, Inés Sánchez and Manuela González for advice regarding statistical analysis. Logistical support was provided by Laboratorio de Ecología Molecular, Estación Biológica de Doñana, CSIC (LEM-EBD). We thank the members of the Conservation and Evolutionary Genetics Group at EBD for constructive comments. C.B.-F. was supported by the University of Panama and Fundación Carolina.

Author contributions

J.A.L. conceived the experiment, C.B.-F. did the lab work under C.R. and J.A.L. supervision, M.C.-S. and C.V. did the statistics, J.A.L. wrote the manuscript with C.B.-F., M.C.-S. and C.V.A. All authors contributed preparation of the final draft, and approved it (J.A.L., C.B.-F., M.C.-S., C.V.A., C.R. and J.L.V.-P.). C.B.-F. and M.C.-S. contributed equally.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Brandariz-Fontes, C. *et al.* Effect of the enzyme and PCR conditions on the quality of high-throughput DNA sequencing results. *Sci. Rep.* **5**, 8056; DOI:10.1038/srep08056 (2015).

Chapter 2 Interglacial refugia on tropical mountains: novel insights from the summit rat (*Rattus baluensis*), a Borneo mountain endemic

Miguel Camacho Sanchez[1], Irene Quintanilla[1], Melissa T. R. Hawkins[2,3*], Fred Y. Y. Tuh[4], Konstans Wells[5], Jesus E. Maldonado[2] and Jennifer A. Leonard[1]

[1]Conservation and Evolutionary Genetics Group, Estación Biológica de Doñana (EBD-CSIC), Avd. Américo Vespucio 26, 41092 Seville, Spain.

[2] Smithsonian Conservation Biology Institute, Center for Conservation Genomics, National Zoological Park, Washington DC 20008, USA.

[3]Division of Mammals, National Museum of Natural History, Smithsonian Institution, P.O. Box 37012, Washington DC 20013-7012, USA.

[4]Sabah Parks, Lot 45 & 46, Tingkat 1-5, Block H, KK Times Square, Coastal Highway, 88100 Kota Kinabalu, Sabah, Malaysia

[5]Environmental Futures Research Institute, School of Environment, Griffith University, Brisbane Qld 4111, Australia

[*]current address: Henry Doorly Zoo and Aquarium, Center for Conservation Genetics, Omaha, NE 68107, USA.

Abstract

Despite much work has been done on post-glacial expansions from refugia, the genetics of organisms currently isolated in interglacial refugia has had much less attention. We performed a field survey of the summit rat (*Rattus baluensis*), and sequenced whole mitochondrial genomes and 27 nuclear markers to test past and present population demographic and connectivity scenarios based on models of habitat distribution through the late Quaternary using Approximate Bayesian Computation. We applied predictions of climate change to forecast the effect of global warming on the conservation of these populations. We found summit rats to be tightly associated with high altitude scrubland habitat, facilitating the modeling of their past, present and future distributions. Our approximate Bayesian analyses on the genetic data support a Holocene fragmentation of a larger population into smaller populations that are now genetically isolated in interglacial refugia on mountaintops, as the upland forest in Borneo retreated to higher elevations following global warming after the Last Glacial Maximum (LGM), ~21 Kya. The current trend of global climate warming will likely lead to diminishing suitable upland habitat and result in the extinction of the Mt. Tambuyukon population by the end of this century. Nevertheless, the population on Mt. Kinabalu could persist at higher elevations, thus highlighting the singular value of high tropical mountains as reservoirs of biodiversity during climate change.

Introduction

The Quaternary is characterized by pronounced environmental fluctuations which have driven changes in the ranges of many species. These dynamics have been closely studied in temperate habitats to explain common phylogeographic patterns and genetic processes in species which experience range contractions during cold periods to southern glacial refugia and then range expansions during interglacials (Hewitt, 1996, 1999, 2000; Petit *et al.,* 2003; Stewart *et al.,* 2010). The opposite pattern, refugia during interglacials, is also possible in other habitats (Stewart *et al.,* 2010), including on tropical mountains (Hewitt, 2000). This lesser studied scenario, referred to as "interglacial refugia" (Bennett & Provan, 2008; Stewart *et al.,* 2010), has been recognized in a handful of studies on tropical mountains including beetles from Central America (MacVean & Schuster, 1981), plants in Africa (Kebede *et al.,* 2007), and mammals from southern China (He & Jiang, 2014; He *et al.,* 2016). However, there is still a significant lack of knowledge on the phylogeographic patterns and genetic processes in these interglacial refugia. Phylogeographic patterns in current populations inform connectivity and their characterization may help identify the opportunities or limitations a species may have in terms of shifting their range in response to future climate change. Tropical mountains offer the opportunity to study these processes as organisms may shift their ranges to higher elevations with increasing temperatures.

At 4,095 m, Mt. Kinabalu, Borneo, is the highest peak in Southeast Asia between the Himalayas and New Guinea, and the most studied mountain in the region. It is protected within Kinabalu National Park, which was declared a UNESCO World Heritage Site in 2000 for its great diversity and endemism. Much of this diversity is associated with its montane habitats, where high levels of endemism are found in plants (Beaman, 2005; Raes *et al.,* 2009; van der Ent *et al.,* 2015), birds (Smythies, 1964), and mammals (Phillipps & Phillipps, 2016). These upper vegetation levels include mossy forest from around 2,000 m, followed by dwarf forest, mountain scrubland and alpine vegetation at around 3,700 m (Kitayama 1992). Most of Mt. Kinabalu's endemics are younger than the formation of the mountain itself, which formed around 6 million years ago (Mya), and originated from speciation from lowland taxa and/or colonization from distant taxa pre-adapted to cool conditions (Barkman & Simpson, 2001; Merckx *et al.,* 2015). During the Last Glacial Maximum (LGM), ~ 21 thousand years ago (Kya), the top of Mt. Kinabalu was covered by glaciers (Stauffer, 1968, 3,658 m snow line; 3,665 m in

Porter, 2001). At this time, the montane forest reached its maximum extension in Sundaland, and ever since has been in regression to higher elevations (Cannon *et al.,* 2009). As montane habitat became reduced and fragmented, it is likely that the species associated with this habitat became isolated in mountaintop refugia.

Such seems to be the case for the summit rat (*Rattus baluensis* Thomas, 1894), a high-elevation endemic which seems to now be in a refugial state. It was previously only known from the upper slopes of Mt. Kinabalu (6.07º N 116.56º E) above 2,100 m (Phillipps & Phillipps, 2016), but here we report our discovery of a second population on Mt. Tambuyukon (2,579 m) (6.20º N 116.66º E; Figure 2.1 A), a peak 18 km away in the same range but with a disjunct upper montane forest isolated by lower elevation forest. We use this system to characterize the genetic processes and degree of connectivity between populations of a species in fragmented and isolated interglacial refugia. We trapped summit rats along an altitudinal gradient on Mt. Kinabalu and Mt. Tambuyukon and sequenced complete mitochondrial genomes and a novel panel of nuclear markers. Using an approximate Bayesian framework, we show an ancestral population was reduced and fragmented in the Holocene, leading to complete genetic isolation. With ongoing climate change we predict an upward shift of suitable habitat of around 500 m by the end of this century. Mt. Kinabalu is the highest of the two peaks and its summit rat population could persist on its upper slopes, but the recently discovered population on Mt Tambuyukon will likely go extinct as the mountain habitat on this lower peak contracts.

Methods

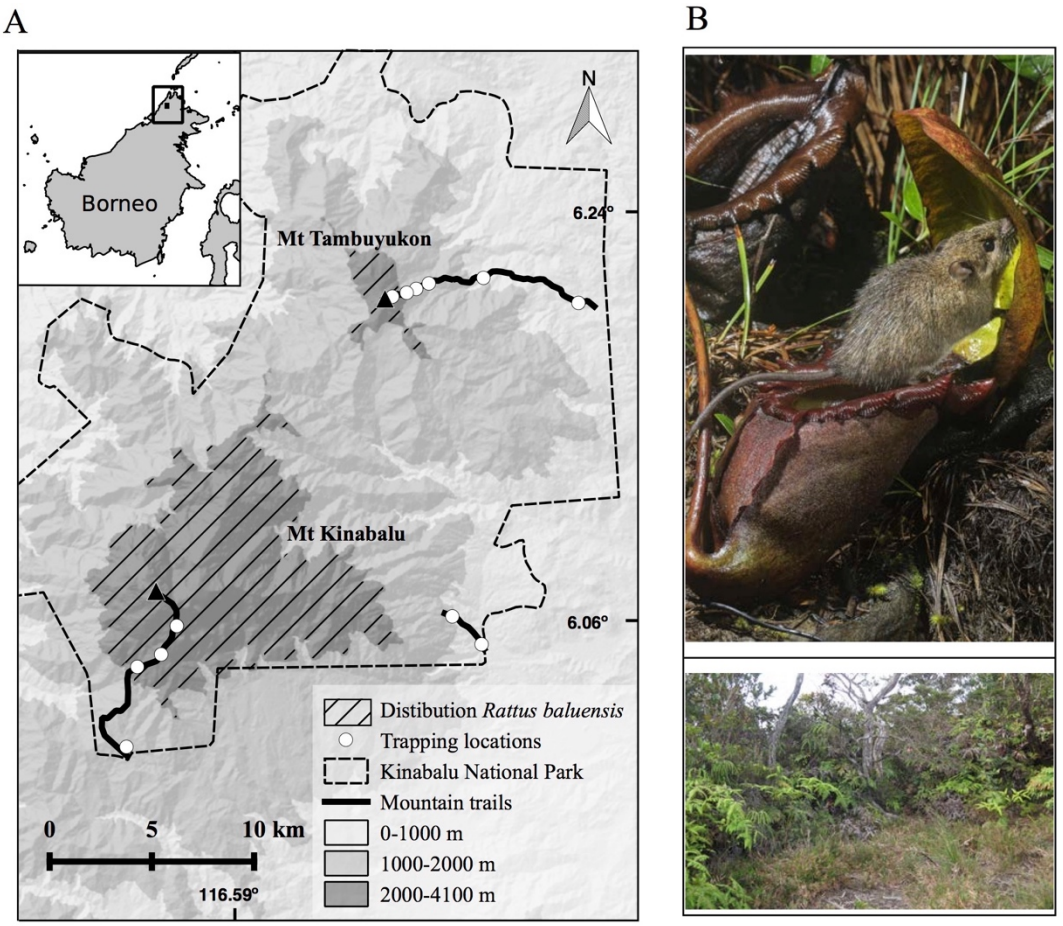
Study system

The summit rat is apparently present continuously across the high mountain habitat (Figure 2.1 B). Its lower distribution limit is situated in the mossy forest just above 2,000 m in elevation (Figure 2.1 A). On Mt. Tambuyukon we recorded it as low as 2,040 m (Table 2.1) and the lowest records from Mt. Kinabalu are around 2,100 (Musser, 1986; Philipps & Philipps, 2016). It has been also observed with camera traps just over 2,000 m in association with pitcher plants (Greenwood *et al.,* 2011; Wells *et al.,* 2011a,b). At these lower elevations it is scarce, but becomes much more abundant in upper mountain dwarf forest and scrubland, up to the summit on Mt. Tambuyukon, and up to at least 3,200–3,426 m on Mt. Kinabalu (Figure 2.1 C; Musser, 1986; Nor, 2001;

Phillipps & Phillipps, 2016). Despite the lack of abundance data above this elevation, there are museum records up to 3,810 m on Mt. Kinabalu (Musser, 1986), and it likely even reaches the summit (Park staff, personal communication). The summit rat has a mutualistic relationship with *Nepenthes rajah*, a pitcher plant found on mountain ultramafic outcrops: the rat climbs on to the pitchers to lick nectar on their lids and the pitcher collects its droppings (Figure 2.1 B). In this way the rat gets sugar in exchange for the nitrogen in its feces (Greenwood *et al.,* 2011; Wells *et al.,* 2011a). The IUCN Red List classifies the status of the summit rat as "Least Concern" due to a presumably large population size and a distribution within a well-protected area (Aplin, 2016).

Field work

We trapped small mammals including the summit rat during two field expeditions to Kinabalu National Park (Sabah, Malaysia) in 2012 and 2013, across two different peaks, Mt. Kinabalu and Mt. Tambuyukon (Figure 2.1 A). Box style live traps were set on altitudinal transects from 331 to 2,509 m following the mountain trail on Mt. Tambuyukon for a total effort of 5,957 trap-nights. On Mt. Kinabalu we trapped for 2,022 trap-nights, from 503 to 1,007 m in and around Poring Hot Springs and from 1,512 to 3,466 m following the Kinabalu summit trail (Fig 2.1; Hawkins 2015). All elevations were extracted using field GPS coordinates on a 1 arc-second SRTM digital elevation model (http://earthexplorer.usgs.gov/). Field samples were collected according to the guidelines of the American Society of Mammalogists (Sikes *et al.,* 2011), as approved by institutional animal care and use committees (Estación Biológica de Doñana Proposal Number CGL2010-21524 and Smithsonian Institution, National Museum of Natural History, Proposal Number 2012-04), with permission from Sabah Parks (TS/PTD/5/4 Jld. 45 (33) and TS/PTD/5/4 Jld. 47 (25)) and the Economic Planning Unit (100-24/1/299), and exported with permissions from the Sabah Wildlife Department (JHL.600-3/7 Jld.7/19 and JHL.600-3/7 Jld.8/) and Sabah Biodiversity Council (Ref: TK/PP:8/8Jld.2). Voucher specimens were deposited at the Doñana Biological Station (EBD), Spain, and Sabah Parks, Malaysia.


A
B
Borneo
N
6.24°
Mt Tambuyukon
Mt Kinabalu
6.06°
Distibution Rattus baluensis
Trapping locations
Kinabalu National Park
Mountain trails
0-1000 m
1000-2000 m
2000-4100 m
0
5
10 km
116.59°

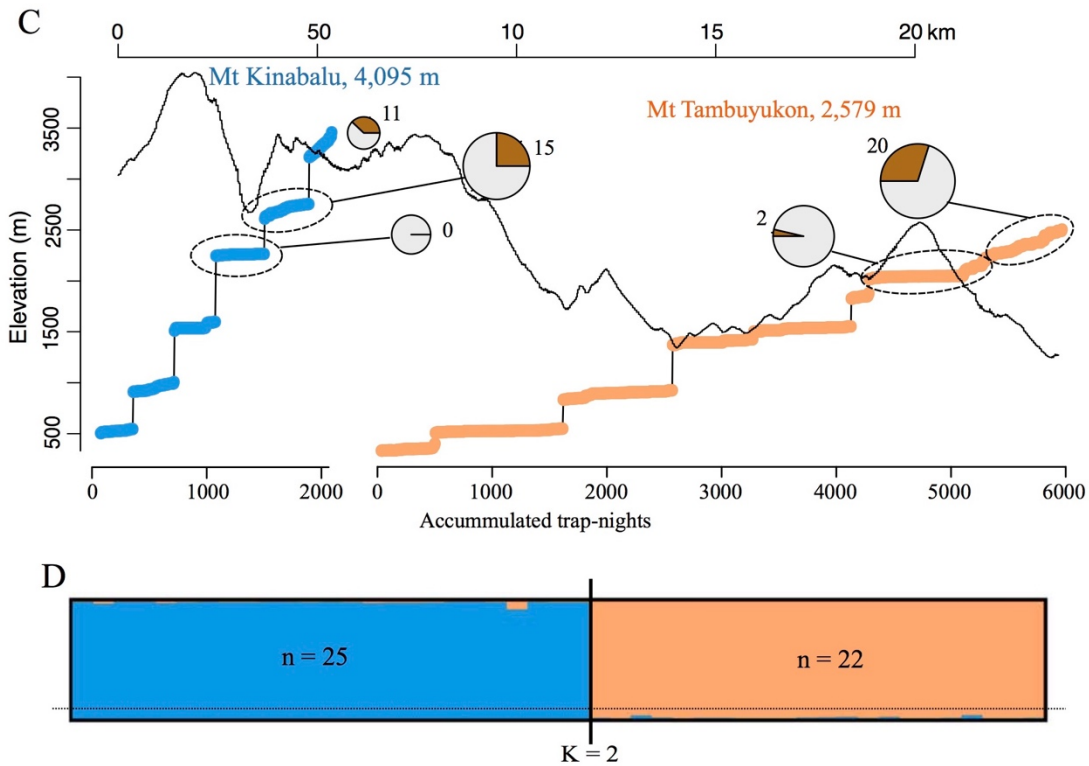

C
0
50
10
15
20 km
Mt Kinabalu, 4,095 m
Mt Tambuyukon, 2,579 m
11
15
0
2
20
Elevation (m)
3500
2500
1500
500
0
1000
2000
3000
4000
5000
6000
Accummulated trap-nights
D
n = 25
n = 22
K = 2

Figure 2-1. A) Study area with trapping locations (open circles) and approximate distribution of the summit rat (hashed area) in Kinabalu Park (marked with dashed line). B) Top: Summit rat licking nectar from the lid of a *Nepenthes rajah* pitcher (photo: Ch'ien C. Lee). Bottom: mountain scrubland habitat. C) Elevation profile across Mt. Kinabalu and Mt. Tambuyukon, with trapping effort across elevation (bottom scale) and number of summit rats trapped (shaded area of pie charts), relative to catches of other small mammals (light area pie charts) at each trapping location. The trapping locations below 2,000 m, where summit rats were not caught, are not depicted. D) Ancestry of the samples from Mt. Kinabalu, left side, and Mt. Tambuyukon, right side, estimated for the most likely number of populations, $K = 2$, with STRUCTURE. The horizontal dotted line indicates a threshold of 0.10 ancestry.

Materials

We included liver and ear samples of 47 summit rats from our field surveys and two additional DNA samples from a previous expedition for a total of 49 samples: 27 from Mt. Kinabalu and 22 from Mt. Tambuyukon (Table 2.1). The tissues we collected were preserved in the field at ambient temperature using NAP buffer (Camacho-Sanchez *et al*., 2013). We extracted the DNA following a standard phenol-chloroform and ethanol precipitation protocol and then quantified with a Nanodrop ND-1000 Spectrophotometer (Nano-Drop Technologies, Inc., Wilmington, DE, USA).

Library preparation and data preprocessing

We amplified a panel of 30 introns, based on the list proposed by Igea *et al.* (2010) to study phylogenies of closely related non-rodent mammal species, although their variation at the intra-species level was unknown. After a test run we selected a panel of 27 primer pairs that amplified target introns in a single reaction optimized to reduce artifacts (Brandariz-Fontes *et al.,* 2015) in the summit rat (Table 2.2). Sequencing primers and individual indices were added in a second pcr. Three individuals were replicated. Genotype calling was done manually, considering only those loci with a minimum of 10x coverage. To ensure reliability in calling heterozygotes, only alleles present in both forward and reverse reads and that had a minimum of 4x coverage were considered (details in Appendix 2.1).

We sequenced complete mitochondrial genomes for all 49 samples (Table 2.1). Dual indexes shotgun or enriched libraries were sequenced on an Illumina MiSeq using 250 bp Paired-End chemistry at the Danish National High-throughput Sequencing Centre (University of Copenhagen, Denmark). Three samples were replicated using the different library preparation protocols to ensure repeatability (details in Appendix 2.1).

Table 2-1. Samples studied. Collector field number, museum code where available, tissue sampled, elevation where animal was caught in meters above sea level, mountain of origin (Location), and average sequence coverage per intron and for complete mitogenomes.

					Coverage	
#Field	**Museum number**	**Tissue**	**Elevation (m)**	**Location**	**intron**	**mitDNA**
BOR326	-	liver	2701	Mt. Kinabalu	84.7	26.9
BOR343	EBD 30374M	liver	2757	Mt. Kinabalu	71.1	44.2
BOR344	-	liver	2757	Mt. Kinabalu	26.1	83.0
BOR348	-	liver	2730	Mt. Kinabalu	73.7	74.2
BOR354	-	liver	2730	Mt. Kinabalu	104.6	43.1
BOR362	EBD 30377M	liver	2757	Mt. Kinabalu	108.9	28.7
BOR364	-	ear	2683	Mt. Kinabalu	87.5	-
BOR372	EBD 30372M	ear	2752	Mt. Kinabalu	64.0	-
BOR373	-	ear	2757	Mt. Kinabalu	74.8	42.0
BOR374	-	ear	2745	Mt. Kinabalu	81.0	-
BOR376	-	ear	2670	Mt. Kinabalu	94.4	-
BOR377	-	ear	2683	Mt. Kinabalu	80.1	-
BOR378	-	ear	2730	Mt. Kinabalu	91.1	-
BOR379	-	ear	2737	Mt. Kinabalu	76.3	-
BOR383	-	liver	3275	Mt. Kinabalu	88.3	81.3
BOR384	EBD 30380M	liver	3294	Mt. Kinabalu	78.9	74.9
BOR391	-	liver	3317	Mt. Kinabalu	106.6	46.5
BOR392	-	liver	3367	Mt. Kinabalu	70.0	19.9
BOR393	EBD 30382M	liver	3336	Mt. Kinabalu	85.7	17.5
BOR398	EBD 30383M	liver	3426	Mt. Kinabalu	166.7	206.1
BOR399	-	ear	3317	Mt. Kinabalu	120.1	115.2
BOR400	-	ear	3235	Mt. Kinabalu	88.6	-
BOR401	-	ear	3275	Mt. Kinabalu	87.2	-
BOR403	-	ear	3382	Mt. Kinabalu	98.8	-
BOR405	-	ear	3222	Mt. Kinabalu	78.3	-
BOR161	-	liver	2050	Mt. Tambuyukon	95.7	37.7
BOR201	-	liver	2449	Mt. Tambuyukon	101.0	37.4
BOR207	EBD 30360M	liver	2377	Mt. Tambuyukon	118.4	140.9
BOR209	-	liver	2363	Mt. Tambuyukon	116.2	96.6
BOR210	EBD 30361M	liver	2381	Mt. Tambuyukon	112.0	26.9
BOR212	-	liver	2477	Mt. Tambuyukon	89.3	36.1
BOR216	-	liver	2363	Mt. Tambuyukon	96.1	28.9
BOR223	-	ear	2449	Mt. Tambuyukon	108.3	93.6
BOR226	-	ear	2363	Mt. Tambuyukon	116.6	-
BOR230	-	liver	2363	Mt. Tambuyukon	119.9	21.0
BOR519	-	liver	2040	Mt. Tambuyukon	134.0	190.5
BOR528	EBD 30395M	liver	2274	Mt. Tambuyukon	101.4	189.2
BOR529	EBD 30396M	liver	2283	Mt. Tambuyukon	81.8	96.9
BOR531	-	ear	2291	Mt. Tambuyukon	85.5	-
BOR532	-	ear	2305	Mt. Tambuyukon	81.2	120.7
BOR533	-	ear	2363	Mt. Tambuyukon	108.7	131.9
BOR534	-	ear	2344	Mt. Tambuyukon	92.7	-
BOR540	EBD 30398M	liver	2194	Mt. Tambuyukon	101.2	197.0
BOR543	-	ear	2291	Mt. Tambuyukon	64.2	-
BOR548	EBD 30400M	liver	2240	Mt. Tambuyukon	175.1	167.0
BOR556	-	ear	2274	Mt. Tambuyukon	61.6	-
BOR557	-	ear	2349	Mt. Tambuyukon	156.7	-
B0993		-	-	Mt. Kinabalu	-	57.5
S0903		-	-	Mt. Kinabalu	-	67.2

Table 2-2. Primers used to amplify the 27 introns and their location in the rat genome. Primer concentration in *μM*, size in base pairs.

Gene name-intron #	Ensembl ID for *Rattus norvegicus* (Rnor_5.0)	Forward primer (5'>3')	Reverse primer (5'>3')	*μM*	Size in summit rat	Location in Rnor_5.0
abcg8-9	ENSRNOG00000005420	TTTCCAATGACTTCCGGGAC	GGCAAAGAAATAAGGACCAGCA	0.65	390	6:7,897,005-7,916,593:1
alkbh7-3	ENSRNOG00000047089	GCTGGAGGTGGCTCTTCTG	CTGGCCTTTCCCTGTTGTCT	1.5	389	9:9,046,025-9,048,157:-1
apeh-17	ENSRNOG00000029572	GAAAGGATGCTGTCTTGGCC	GGGGTGGCCTTGGTTGTATA	0.85	415	8:116216748-116225863:-1
apeh-14	ENSRNOG00000029572	KGACACCCATGACACAGACT	CCCAGTTCTCCACACCCA	-	NA*	8:116216748-116225863:-1
cd27-5	ENSRNOG00000027466	CAGGCTCRGGTTTCCGGT	TCCGGATCTTTGTGACCTTCT	0.75	379	4:224,761,678-224,766,884:-1
chrna9-1	ENSRNOG00000002484	TTATCTGGGAGAGCGTGACC	TTGGGAAARGATGAACCGGC	1.0	408	14:43,670,533-43,677,246:-1
dhrs3-3	ENSRNOG00000015736	CTCCTCAAGTCCCAGCATGT	GCACRGAATTGAGGCACACA	1.25	396	5:166,490,225-166,524,733:1
fancg-9	ENSRNOG00000010424	CCTTTAGTTGTGACCAGGCC	GAGCATTACCTGGACCTGCT	1.0	447	5:62,973,516-62,981,572:-1
fetub-1	ENSRNOG00000001806	ACAGAGAKCCCATGTCTTCC	GCCCTGCAGAACATCAACAG	0.75	393-394	11:84745603-84756917:-1
fn3krp-5	ENSRNOG00000036660	AGATGGACATGGTGGAGAAGA	AGTGKCCATAGAAGGATGCT	1.5	403	10:110199929:110209824:1
gabrp-1	ENSRNOG00000032417	TCTGCTGACCTCCACATTGA	AGCTACAGYCTCTATTTGGCCT	-	NA*	10:18,485,514-18,506,337:-1
gadd45g-1	ENSRNOG00000013090	GACCTCCAAGTCCCAGCTG	GGATACAGTTCCGGAAAGCAC	1.0	402	17:15469301:15471060:1
il34-3	ENSRNOG00000017602	GGTACTCAGAGTGGCCAACA	CCAGCAATGTCTGAACCTCC	0.8	397	19:51,714,803-51,728,622:1
irf5-7	ENSRNOG00000007437	AAACCCCGAGAGAAGAAGCT	CTGGACCATGGGCTGCAA	1.0	385	4:56,572,293-56,583,432:1
klc2-10	ENSRNOG00000020299	AAAGCCCTACCTGTTTGCG	TCAGGATAAGCGCCGGGA	0.75	390	1:227,423,230-227,433,026:-1
mmp9-2	ENSRNOG00000017539	GATGATGGGAGAGAAGCAGTC	GTCTCGCGGCAAGTCTTC	1.75	409-419	3:167597817:167605882:1
ms4a2-5	ENSRNOG00000020993	ACACCAGTTCCTGTCAAACA	CTYCGCTTATATGAACWACTGCA	1.75	370	1:235019784:235027733:1
mycbpap-11	ENSRNOG00000042912	GGCAGAATCACACCTGGGA	GGTCAATAACGGCACKGTGG	1.35	416	10:82,051,977-82,072,661
nfkbia-5	ENSRNOG00000007390	GCCTCCAAACACACAGTCAT	TGAGGAGAGCTATGACACGG	1.25	482-486	6:85803391:85807823:-1
npr2-10	ENSRNOG00000015991	TGAACTCAAACACGTACGTACT	TGGTTGAACTGRACATCTCTCA	1.75	424-428	5:63,653,695-63,672,094:1
p2rx1-3	ENSRNOG00000017606	CATTGTGCAGAGGTGAGGAC	TCTGCTTTTCCTGGAGTGCA	1.5	420-421	10:59,305,745-59,320,794:1
pipox-5	ENSRNOG00000008798	TCTGAGAAGGTTTTGGGGCA	CCCACCACATCTAYGGACTG	0.55	390	10:66,660,854-66,673,403:-1
ptgs2-7	ENSRNOG00000002525	GTGTATCCYCCCACAGTCAAA	TGAGTTTGAAGTGGTAACCGC	1.0	422	13:72,316,370-72,323,056:1
rabac1-1	ENSRNOG00000020233	AATACTCCACGTTGCGWACC	CAGAAGGACCAGCAGAAGGA	1.25	411	1:83095962:83099071:1
rgd735029-5	ENSRNOG00000019276	CTTCGGAGGCATGTTCTTCC	CCTTTGCCTGGGATGYGAAG	1.5	374-376	18:31,076,952-31,090,582:1
rogdi-7	ENSRNOG00000003125	AGAARCCGGCTCACTACCC	GAGGCACAGCTTGTTGAGG	1.0	394-398	10:9,528,490-9,533,108:1
sfrs5-1	ENSRNOG00000005513	TCAAGGGTTACGGACGGATC	TCATCTGCATCCCTTGGGTC	-	NA*	6:104,611,026-104,615,302:1
ssfa2-13	ENSRNOG00000005865	ACCCTCATATGACAGAGGAGG	ATTCGGACAGAGTTCCGCA	1.25	383	3:73,152,918-73,190,192:1
tmem87a-16	ENSRNOG00000008455	CTGCTTGGTACTTCTCATTTTCA	TGTCAGAGGAAGATGARGAGGA	1.75	400-403	3:118,677,063-118,721,810:-1
usp20-17	ENSRNOG00000007710	AACGTGATCAATGGGCAGTG	AGGAAGGTGTGGTTGGTGAT	0.85	403	3:15,181,259-15,214,732:1

*NA: it failed to amplify in the multiplex PCR.

Nuclear genetic structure and diversity

We assessed population structure with a Bayesian clustering method implemented in STRUCTURE v 2.3.4 (Pritchard *et al.,* 2000). We used an admixture model using locations as prior and assuming correlation of allele frequencies among populations. For each *K* from 1 to 6, we performed four independent runs with 50,000 MCMC repetitions and a burn-in period of 10,000. We generated consensus solutions for each *K* (1-6) by clustering the four runs with CLUMPAK (Kopelman *et al.,* 2015). Then, STRUCTURE HARVESTER (http://taylor0.biology.ucla.edu/structureHarvester, Earl & VonHoldt, 2012) was used to infer the most probable number of populations through the ΔK method (Evanno *et al.,* 2005).

The number of alleles, the observed and expected heterozygosity, and the fixation index for each locality were calculated in GenAlEx 6.5 (Peakall & Smouse, 2012). We also evaluated the partitioning of nuclear genetic variation with an AMOVA and the genetic differentiation, F*st*, in GenALEx. We calculated the nucleotide and haplotype diversity of introns with DNAsp version 5.10.1 (Librado & Rozas, 2009) with indels coded as a fifth state.

To visualize the haplotype diversity in the introns, we aligned the alleles in all the samples for each of the 19 polymorphic loci. We considered indels as a fifth character and used these alignments to build TCS haplotype networks (Clement *et al.,* 2000) in PopART (http://popart.otago.ac.nz).

Mitochondrial structure and diversity

The complete mitogenomes of the 32 samples (16 from Mt. Kinabalu and 16 from Mt. Tambuyukon) were aligned with the MAFFT v7.017 (Katoh & Standley, 2013) plugin in Geneious 8.1.5. The mitochondrial matrix only had 0.01 % missing data. We used the alignment to build a TCS haplotype network in PopART (http://popart.otago.ac.nz). Most population genetics and phylogeography studies on rodents use either *cytochrome b* or the control region alone. We also built haplotype networks for these regions extracted from the complete mitogenomes in PopART to compare the level of information provided by complete mitogenomes to these more commonly utilized mitochondrial markers.

We used DNAsp v5.10.01 to calculate the number of polymorphic sites, parsimony-informative sites, haplotype diversity, number of haplotypes, nucleotide diversity (π),

and Tajima's D for the complete mitogenome alignment, and for the *cytochrome b* and control region separately. We assessed the partition of the molecular variance between and within populations with an AMOVA in GenALEx 6.5, in which each polymorphic position was considered a different locus. The complete mitochondrial alignment had a total of 16,315 positions. Indels and missing data are not considered in PopART haplotype networks or DNAsp, and so were removed in GenALEx, yielding a matrix of 16,264 positions of which 157 were polymorphic.

Demographic history

We used an Approximate Bayesian Computation (ABC) approach in PopABC 1.0 (Lopes *et al.,* 2009; https://sites.google.com/site/jsollarilopes/) using sequences from the polymorphic introns to estimate the time when the Mt. Kinabalu and Mt. Tambuyukon populations split. PopABC samples parameters from given prior distributions, which is used to simulate data from which summary statistics are calculated. A rejection step then discards all but a given proportion of the simulations whose statistics are closest to the summary statistics of the real data. The set of priors included the time of the split between the Mt. Kinabalu and Mt. Tambuyukon populations (*tev*), the effective population size of Mt. Kinabalu (*Ne1*), the effective population size of Mt. Tambuyukon (*Ne2*), the effective size of the ancestral population (*NeA*), and a mutation rate (μ). We explored a range of priors sampled from uniform distributions (Table 2.3). The range for time of split (*tev*) was set from 1 to 20 Kya, coinciding with the LGM. For defining upper prior limits for populations sizes, we classified all area above 2,000 m from a SRTM digital elevation model (http://earthexplorer.usgs.gov/) in QGis 2.8 (QGIS Development Team 2015) as potentially suitable habitat. We did not have estimates of population densities for this species. We explored the upper limits for the *Ne* priors using density estimates from the invasive rat *Rattus exulans* available in PanTHERIA (Jones *et al.,* 2009). This rat has the largest densities for the genus *Rattus* reported in PanTHERIA, 7155 rats/km^2. Effective population sizes in mammals are often 10% of census size, but to be even more generous in estimating the maximum we used a factor of 0.2 to convert the estimated maximum census size to the estimated maximum effective population size (Allendorf & Luikart, 2007). After exploring simulations in PopABC the priors for *Ne1* and *Ne2* were set to 10 - 100,000 individuals for Mt. Kinabalu and 10 - 25,000 individuals for Mt. Tambuyukon (Table 2.3). We set a larger upper limit for *NeA*, 200,000 individuals, since we hypothesize that before the split the

suitable habitat for the summit rat was larger. We calculated the mutation rate of the 27 introns using Bayesian inference in BEAST 1.8.0 (Drummond *et al.,* 2012). To determine the best model of sequence evolution we aligned one of the alleles for each the 27 loci from the summit rat with homologous sequence from *Mus musculus* and *Rattus norvegicus* (the same sequences used for primer design). We used Gblocks 0.91b (Castresana, 2000) to remove poorly aligned regions using default parameters, except minimum length of a block set to 3. Then, with Gblocks we generated a concatenated alignment with all 27 introns, which we used as input to determine the best model of sequence evolution in JModelTest.2.1.7 (Darriba *et al.,* 2012). According the BIC criterion, the best model of sequence evolution was HKY. The analysis in BEAST assumed uncorrelated lognormal relaxed clocks for each intron, a HKY substitution model with estimated base frequencies, and a Yule model of speciation. We constrained the *Rattus*-*Mus* split to a normal distribution with a mean of 11.81Mya and *s.d*. of 0.4 My (Kimura *et al.,* 2015). We ran three chains for 50 million generations sampled every 10,000 generations. The convergence of the runs was confirmed in Tracer v1.6.0. The Effective Sample Size (ESS) was above 200 in each case. Also in Tracer, we extracted the mutation rates for each intron as the geometric mean after discarding at least the first 10% of the generations for each run. PopABC requires the mutation rate to be in mutations per generation per locus. We used the generation time inferred for the summit rat in Pacifici *et al.* (2013): 535 days (1.47 years), which is the addition of the age at first reproduction plus a term which summarizes the reproductive life span and the relative fecundity of the young versus the old individuals in the population. Indels were removed from the input data matrix, as the mutation rates had been initially estimated for substitutions alone. We discarded locus *pr2x1*-3 because its variation was due to indels alone. So, a total of 18 polymorphic loci were kept. Their average size without indels was 401 bases, and their average mutation rate was calculated to be 4.37×10^{-6} substitutions per generation per locus.

We ran 500,000 simulations in PopABC with priors as in Table 2.3. A rejection step was run to keep only the closest 1% simulations to the real data according to the nine statistics for sequence data that PopABC computes. A PCA in R confirmed the accepted simulations were close to the study data (Appendix 2.2). Then, we ran a regression following the strategy as in Beaumont *et al.* (2002), by using the R scripts *make_pd2.r*, *loc2plot_d.r* and *reg_step.r*, distributed with PopABC 1.0 in R 3.1.3 (R Core Team,

2015). In the regression step, each accepted set of parameters is given a weight according to its distance to the real data. We plotted the fitted data following *reg_step.r* and calculated the mode and the 95 % highest posterior density (HPD) intervals of the posterior distribution of the parameters with function *loc1statsx* in *loc2plot_d.r*.

Table 2-3. Prior and posterior distributions of the parameters simulated with PopABC.

			Posterior distribution	
Parameter	Description	Prior	mode	95 % HPD
tev	Years from split	uniform [1, 20,000]	2,005	546 – 4,529
Ne1	Population size on Mt. Kinabalu	uniform [1, 100,000]	22,891	7,102 – 61,142
Ne2	Population size on Mt. Tambuyukon	uniform [1, 25,000]	4,290	1,178 – 9,171
NeA	Ancestral effective population size	uniform [1, 200,000]	40,534	20,448 – 66,206
μ	Average mutation rate	$4.46*10^{-6}$	-	-
μ_{sd}	Std. dev. of μ across loci	$1.74*10^{-6}$	-	-

Results

Trapping and ecology

Although the summit rat was considered endemic to Mt. Kinabalu (Phillipps & Phillipps, 2016), we discovered a second population on Mt. Tambuyukon, another peak in the same mountain range, 18 km north (Figure 2.1 A). We recorded summit rats on transects that ranged in elevation from 2,040 to 2,477 m on Mt. Tambuyukon and from 2,670 to 3,426 m on Mt. Kinabalu (Table 2.1) in a total of 2,276 trap-nights. At these locations we trapped a total of 211 different small mammals of which 48 were summit rats (Figure 2.1 C), with an overall trapping success ranging from 5.6 % to 15.4 % depending on the transect. Despite a larger trapping effort of 5,703 trap-nights, the summit rat was never trapped in transects at lower elevations: 2,250-2,268 m and below on Mt. Kinabalu and 1,504-1,881 m and below on Mt. Tambuyukon (Figure 2.1 C). On Mt. Tambuyukon, only two summit rats were trapped in mossy forest (2022 m – ~2290 m), but it was relatively common (~1/3 of the catches) in the mountain dwarf forest and scrubland (~2290 m – 2,509 m), particularly in areas with the pitcher plant *Nepenthes rajah* (Appendix 2.3). On Mt. Kinabalu, we did not trap any summit rat in the mossy forest at the 2,250-2,268 m transect, but it was again relatively abundant in upper dwarf forest and scrubland in the two upper transects: 2,615-2,759 m, ~1/4 of the catches, and at 3,222-3,466 m (~1/3 of the catches).

Sequencing

Between the two Roche 454 GS Junior runs a total of 126,825 reads more than 250 bp long were assigned to one of the 47 animals sequenced (not considering replicates). Most of those reads, 121,636 in total, mapped to one of the 27 introns in *Rattus norvegicus*, yielding an average coverage of 96 reads per locus per individual (Table 2.1). The genotypes for all three replicates were consistent, although some allelic dropout was observed in cases of low coverage. The intron sequences are in GenBank under accession numbers xxxx – xxxx (not yet submitted).

Whole mitochondrial genomes (≥ 99.5% completeness) were reconstructed from 32 of the 49 samples attempted (GenBank KY611359 - KY611390) with an average coverage of 82.5x (Table 2.1). We found no mismatches in any of the replicated samples, although there were up to two ambiguous positions.

Nuclear genetic diversity and structure

We amplified and sequenced the 27 targeted introns. One of them, *mycbpap*-11, was discarded because of low coverage. Of the remaining 26 introns, 6 were monomorphic (*alkbh7*-3, *apeh*-17, *fancg*-9, *irf5*-7, *pipox*-5 and *ptgs2*-7) and were thus discarded. One of the polymorphic introns (*fetub*-1) had a very high observed heterozygosity (Mt. Kinabalu: 0.84; Mt. Tambuyukon: 0.91) despite a much lower expected heterozygosity (Mt. Kinabalu: 0.49; Mt. Tambuyukon: 0.51). There were two alleles of similar size (393 and 394 bp) but a high number of polymorphic sites (S = 22). These results suggest that they may be different loci instead of different alleles and were consequently discarded for downstream analysis. The remaining 19 loci were used to assess the nuclear genetic diversity and the structure of the populations.

The polymorphic introns had from two to five alleles per locus: nine introns had two alleles, five introns had three alleles, four introns had four alleles and one had five alleles (Table 2.4; Figure 2.2). Fifteen out of 50 alleles were private to Mt. Kinabalu and five out of 38 were private to Mt. Tambuyukon. The average nucleotide and haplotype diversities were also higher on Mt. Kinabalu (π = 0.00163; Hd = 0.31) than on Mt. Tambuyukon (π = 0.00119; Hd = 0.23) (Table 2.4). The mean observed heterozygosity was higher on Mt. Kinabalu (mean ± sd: 0.24 ± 0.05) than in Mt. Tambuyukon (mean ± sd: 0.17 ± 0.04) (Table 2.4). The fixation index (F), a measure of the inbreeding

coefficient, was not different from zero in either population (mean ± sd; Mt. Kinabalu: -0.05 ± 0.02; Mt. Tambuyukon: -0.05 ± 0.06).

Table 2-4. Population genetic statistics on summit rats from each mountain separately and all together. Number of polymorphic sites (S), number of haplotypes (h), haplotype diversity (θ), nucleotide diversity (π x10-3), observed heterozygosity (Ho), expected heterozygosity (He) and fixation index (F) for the 19 polymorphic nuclear loci.

	Kinabalu							**Tambuyukon**							**Overall**			
Locus	S	h	θ	π	Ho	He	F	S	h	θ	π	Ho	He	F	S	h	θ	π
abcg8-9	1	2	0.04	0.10	0.04	0.04	-0.02	0	1	0.00	0.00	0.00	0.00	-	1	2	0.02	0.05
cd27-5	4	5	0.55	2.57	0.56	0.53	-0.05	2	2	0.17	0.89	0.18	0.17	-0.10	4	5	0.40	1.94
chrna9-1	3	2	0.49	3.60	0.56	0.48	-0.17	3	2	0.21	1.69	0.23	0.20	-0.13	3	2	0.47	3.47
dhrs3-3	2	3	0.15	0.40	0.16	0.15	-0.06	1	2	0.05	0.11	0.05	0.04	-0.02	2	3	0.10	0.26
fn3krp-5	1	2	0.30	0.75	0.28	0.30	0.05	1	2	0.44	1.09	0.43	0.43	0.00	1	2	0.37	0.90
gadd45g-1	3	2	0.04	0.31	0.04	0.04	-0.02	3	2	0.05	0.38	0.05	0.04	-0.02	3	2	0.04	0.32
il34-3	1	2	0.08	0.20	0.08	0.08	-0.04	0	1	0.00	0.00	0.00	0.00	-	1	2	0.05	0.11
klc2-10	0	1	0.00	0.00	0.00	0.00	-	1	2	0.51	1.31	0.64	0.50	-0.27	1	2	0.36	0.93
mmp9-2	5	4	0.46	1.76	0.52	0.45	-0.15	4	3	0.60	2.93	0.41	0.58	0.29	5	4	0.52	2.25
ms4a2-5	5	4	0.29	1.64	0.28	0.28	0.01	0	1	0.00	0.00	0.00	0.00	-	5	4	0.16	0.89
nfkbia-5	1	2	0.15	0.31	0.16	0.15	-0.09	1	2	0.13	0.27	0.05	0.13	0.64	2	3	0.14	0.30
npr2-10	5	4	0.65	5.33	0.56	0.63	0.12	4	3	0.53	3.71	0.38	0.52	0.26	5	4	0.61	4.72
p2rx1-3	1	2	0.50	1.19	0.61	0.49	-0.24	1	2	0.05	0.12	0.05	0.05	-0.03	1	2	0.44	1.04
rabac1-1	1	2	0.50	1.22	0.48	0.49	0.03	1	2	0.13	0.32	0.05	0.13	0.64	1	2	0.40	0.96
rgd735029-5	4	4	0.51	1.61	0.44	0.50	0.12	1	2	0.10	0.25	0.10	0.10	-0.05	4	4	0.36	1.09
rogdi-7	7	2	0.33	5.86	0.42	0.33	-0.26	8	3	0.47	4.95	0.41	0.46	0.11	8	3	0.57	9.28
ssfa2-13	0	1	0.00	0.00	0.00	0.00	-	1	2	0.50	1.31	0.59	0.49	-0.20	1	2	0.40	1.03
tmem87a-16	5	3	0.31	2.89	0.28	0.30	0.07	4	2	0.17	1.71	0.19	0.17	-0.11	5	3	0.25	2.35
usp20-17	4	3	0.54	4.52	0.64	0.53	-0.20	4	2	0.36	3.57	0.45	0.35	-0.29	4	3	0.46	4.10
MEAN	2.8	2.6	0.31	1.63	0.24	0.23	-0.05	2.1	2.0	0.23	1.19	0.17	0.17	-0.05	3.0	2.8	0.32	1.88
SD	2.1	1.1	0.21	1.8	0.05	0.05	0.02	2.0	0.6	0.2	1.5	0.04	0.04	0.06	2.0	1.0	0.2	2.2

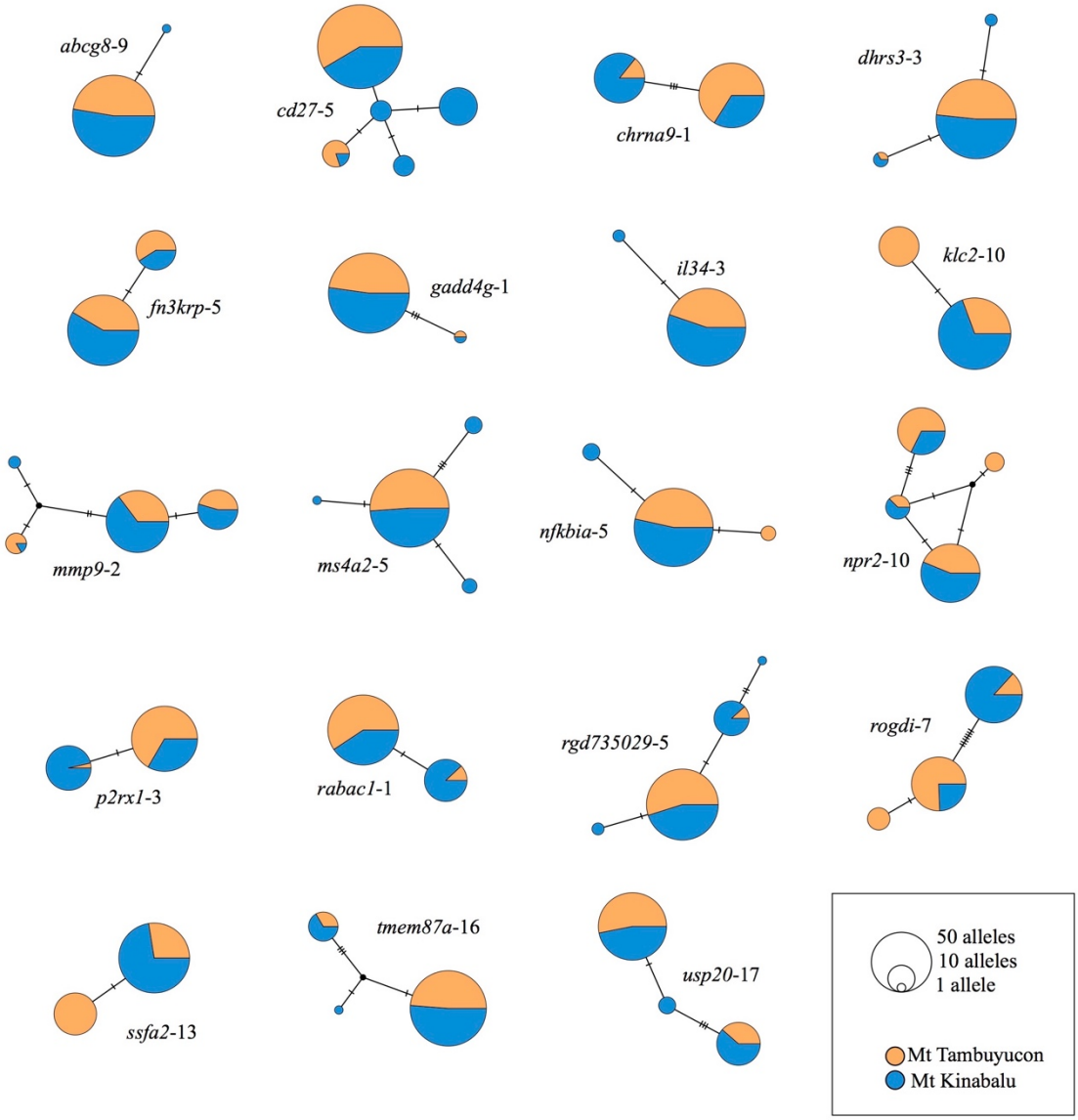


Figure 2-2. TCS haplotype networks for each of the 19 polymorphic introns studied. The size of the pie charts corresponds to the frequency of the haplotype; transversal lines represent single base pair differences or indels; missing haplotypes.

The STRUCTURE analysis revealed that the most likely number of genetic groups was two ($K = 2$), one for each mountain (Figure 2.1 D; Appendix 2.4). All the samples clustered according to the mountain they came from and had nearly 100% ancestry associated with their geographical population. Increasing the number of clusters up to $K = 6$ did not reveal any further genetic structure (Appendix 2.5). An AMOVA indicated that 22 % of the genetic variation was partitioned between the two mountains, whereas the variation within individuals accounted for 68%, and only 10% among individuals within the clusters. The F*st* was 0.22 ($P = 0.01$) indicating a strong genetic differentiation between the rats on the two mountains.

Mitochondrial diversity and structure

There were 12 different haplotypes in the 32 complete mitogenomes. All the haplotypes were unique to one of the two mountains (Figure 2.3). Haplotype diversity was higher on Mt. Kinabalu (Hd = 0.800), with 8 haplotypes, than on Mt. Tambuyukon (Hd = 0.642), with 4 haplotypes (Table 2.5). All the haplotypes differed by few base pairs (1 – 18) except one haplotype found in only one animal from Mt. Tambuyukon, which was 131 base pairs different from the most similar haplotype in the network (Figure 2.3). This divergent haplotype led to increased nucleotide diversity and number of polymorphic sites in Mt. Tambuyukon with respect to Mt. Kinabalu (Table 2.5). When only *cytochrome b* or the control region were considered, the number of haplotypes decreased to 5 and 7, respectively. The amount of information as measured by parsimony informative sites decreased from 22 when considering complete mitogenomes, to 2 with only *cytochrome b* or 5 with the control region (Table 2.5; Appendix 2.6).

An AMOVA assigned 22 % of the genetic variance ($P = 0.001$) to the differences between the two mountains.

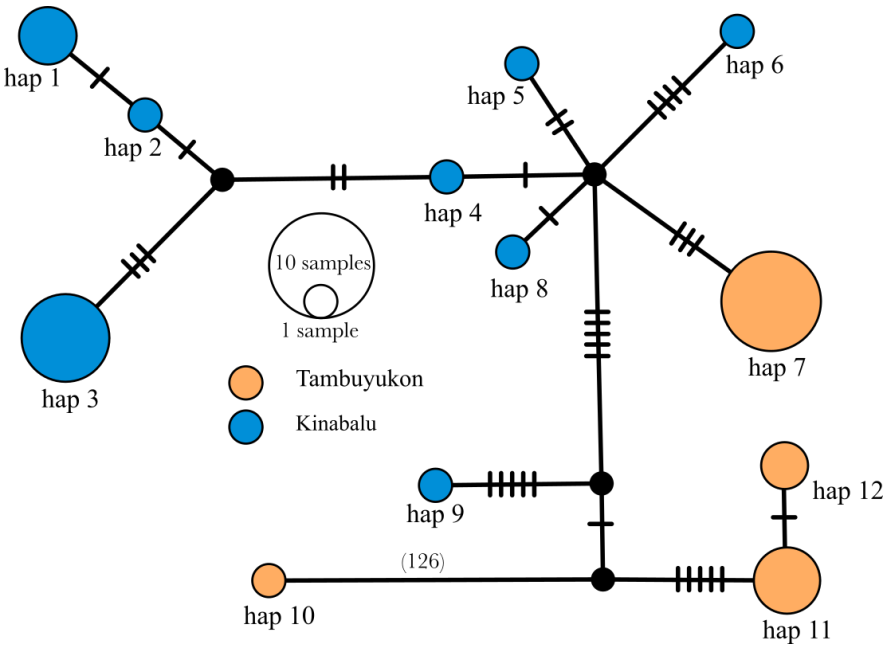


Figure 2-3. TCS haplotype network for the complete mitochondrial genomes; 16 from Mt. Kinabalu and 16 from Mt. Tambuyukon. The pie chart size corresponds to the frequency of each haplotype; transversal lines represent a single base pair differences; small black circles represent missing haplotypes.

Table 2-5. Genetic diversity for complete mitochondrial genomes, *cytochrome b* and control region of 32 samples. Number of polymorphic sites, S, parsimony informative sites, PI, haplotype diversity, θ, number of haplotypes, h, nucleotide diversity, π (10^{-3}). MK, Mount Kinabalu; MT, Mount Tambuyukon (ns: non-significant;*, P < 0.05; **, P < 0.01).

	mitogenomes			*cytochrome b*			control region		
	MK	MT	overall	MK	MT	overall	MK	MT	overall
S	24	140	157	2	13	14	3	17	19
PI	8	15	22	2	1	2	2	3	5
θ	0.800	0.642	0.865	0.717	0.575	0.762	0.725	0.575	0.831
h	8	4	12	4	3	5	4	3	7
π	0.35	1.41	0.101	0.82	1.77	1.53	1.14	3.66	3.23
Tajima's D	-0.91 ns	-1.97*	-2.22**	1.40ns	1.86*	-1.64 ns	0.25 ns	-1.43 ns	-1.40 ns

Demographic history

The posterior distributions of the parameters in the PopABC analyses show that the summit rat populations from Mt. Kinabalu and Mt. Tambuyukon became isolated approximately 2,000 years ago (95% HPD: 546 – 4,529 years) (Table 2.3; Figure 2.4). Before the split, the effective population size of the ancestral population size (*NeA*) was estimated to be around 40,500 (95 % HPD: 20,448 – 66,208). However, the effective population sizes estimated for the current populations are much lower: approximately 4,300 on Mt. Tambuyukon (95 % HPD: 1,178 – 9,171), and larger on Mt. Kinabalu, 23,000 (95 % HPD: 7,102 – 61,142; Table 2.3; Figure 2.4). Assuming the potential distribution of this species as the area above 2,000 m and a factor of 10 between effective population size and census size (Frankham, 1995), the densities of the summit rat on Mt. Kinabalu and Mt. Tambuyucon are 2,100 rats per Km^2 and 5,400 rats per Km^2, respectively. These densities are in the middle of the range of densities reported for other species in the genus *Rattus* in the database PanTHERIA (Jones *et al.,* 2009).

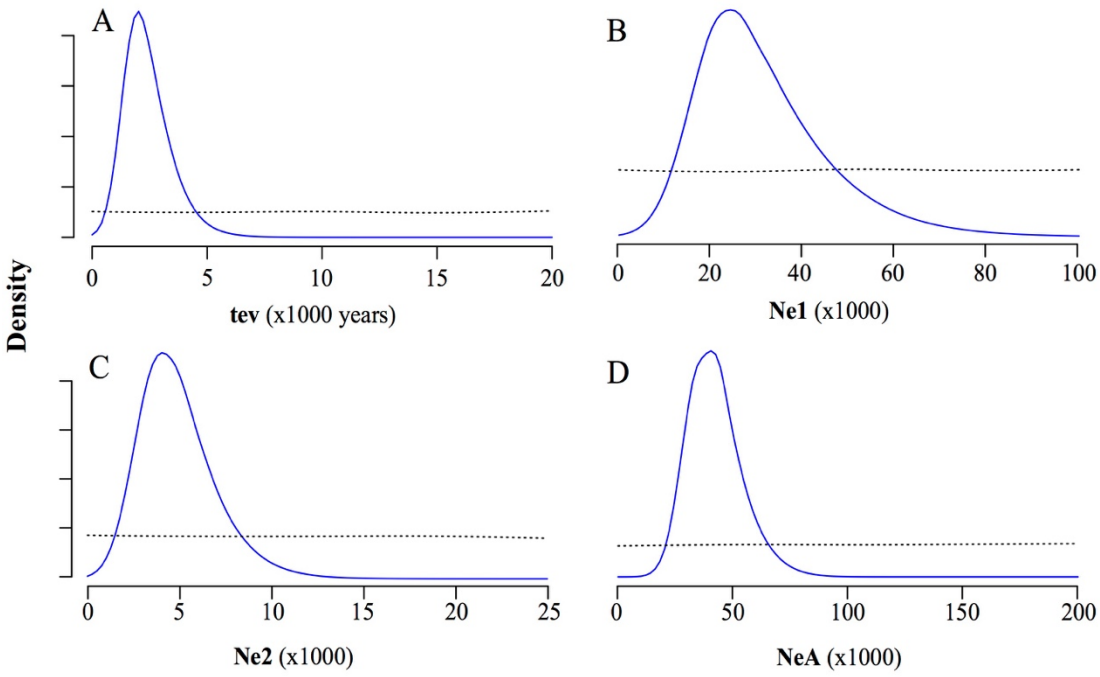


Figure 2-4. Prior (dashed line) and posterior (solid line) distributions for the time of split (*tev*), ancestral effective population size (*NeA*), and effective population size of summit rats on Mt. Kinabalu (*Ne1*) and Mt. Tambuyukon (*Ne2*) used in the PopABC analyses.

Discussion

Distribution and ecology of the summit rat

The summit rat was previously described as endemic to Mt. Kinabalu (Musser, 1986; Phillipps & Phillipps, 2016). We discovered a new population on Mt. Tambuyukon, a peak in the same mountain range 18 km north, where we trapped individuals between 2,040 and 2,477 m. These data, along with previous records (Nor, 2001; Wells et al., 2011b), suggest a lower distribution limit for the summit rat corresponding to the lower delineation of the cloud forest (mossy forest or upper montane forest) at ~2,000 m (Kitayama, 1992), although it is at very low densities in this habitat. The genetic data supported a strong genetic isolation of the populations on Mt. Kinabalu and Mt. Tambuyukon. These genetic data support the absence of summit rats in lower altitude forest that connects the two mountains (at around 1700 m) and may suggest that the lower mountain forest is a barrier to dispersal between these two populations.

We found summit rats to be most abundant in the higher altitude dwarf forest and montane scrubland above ~2,300 m and up to at least 3,426 m (Figure 2.1 C) as

previously reported (Musser, 1986; Nor, 2001; Phillipps & Phillipps, 2016). There is no abundance data above that elevation, just occasional records up to 3,810 m on Mt. Kinabalu (Musser, 1986). Densities are probably lower at these elevations given the low productivity in these subalpine and alpine zones in Mt. Kinabalu (Kitayama, 1992). This hypothesis is supported by the estimated population densities for Mt. Kinabalu, 2,100 rats/Km^2, which was about half compared to that estimated for Mt. Tambuyukon, 5,400 rats/Km^2. This lower density for Mt. Kinabalu likely reflects averaging across the more productive scrubland where densities could be similar to Mt. Tambuyukon, and the much less productive alpine and subalpine habitats where densities are likely much lower (Kitayama, 1992). These alpine habitats are absent on Mt. Tambuyukon because of its lower elevation as compared to Mt. Kinabalu.

The summit rat seems to be particularly abundant in areas with presence of *Nepenthes rajah* (Appendix. 2.3). The summit rats have previously been reported to have a mutualistic relationship with *Nepenthes rajah*, a pitcher plant that is also a Kinabalu Park endemic (Greenwood et al., 2011; Wells et al., 2011a). The summit rat feeds on the nectar provided by pitcher plants mainly at night, while the mountain treeshrew (*Tupaia montana*) feeds on the nectar during the day. Given the high rate of visits to the plants by the two species, these pitcher plants may represent an important food source in the areas where they are sympatric (Greenwood et al., 2011; Wells et al., 2011a). In turn, the feces of the rats and treeshrews may be an important source of nitrogen for *Nepenthes rajah* in the impoverished ultramafic soils it is restricted to (Clarke et al., 2009; Van der Ent et al., 2015). However, the distribution of *Nepenthes rajah* is patchy, and *Rattus baluensis* is also very abundant at areas without its presence. *Tupaia montana* also spans down areas 900 m elevation. This suggests that the mutualism is not dependent at least on the mammal's side.

Effects of late Quaternary changes

We found strong genetic isolation between the populations of summit rats on Mt. Tambuyukon and Mt. Kinabalu that probably derives from a Holocene fragmentation of these two populations from a larger widespread ancestral one incorporating both mountain tops in a larger ancestral distribution area (Appendix 2.7). These findings are consistent with models that predict Late Quaternary retreat of the upland vegetation to higher elevations in parallel with temperature increase after the LGM. Recent models predict an increase of 3ºC in temperature and a vegetation lapse rate of 166 m /ΔºC for

the upland forest since the LGM in Sundaland, suggesting that habitats have shifted up mountains by approximately 500 m (Cannon *et al.,* 2009; Struebig *et al*., 2015). Specific modeling of the cloud formation upon which the cloud forest on Mt. Kinabalu depends predicted an increase in altitude of 439 m since the LGM (Still *et al.,* 1999). Geological and palynological evidence reveal an even more severe temperature change in the mountains of tropical Southeast Asia since the LGM. On Mt. Kinabalu, a glaciated cap reached around 3,660 m, which implies depression of the snowline by around 990 m, equivalent to around Δ 5.4 ºC from the LGM to recent (Porter, 2001). Similar evidence from glaciers in New Guinea, which also suggest a temperature increase of 5-6 ºC on their mountaintops from the LGM, have been reported (Hope, 2007). These values are also similar to the 6–9 ºC increase that pollen data suggest for mountains in tropical southeast Asia and Australia since the LGM (Pickett *et al.,* 2004). At the LGM, summit rats could have expanded their distribution to much lower elevations of around 720 m assuming an increase of 6 ºC on Mt. Tambuyukon and Mt. Kinabalu since the LGM, a lapse rate of 5 ºC/Km for tropical mountains (Sekercioglu *et al.,* 2008), and a correction of -120 m due to sea level change (Cannon *et al.,* 2009). If this estimation is accurate, it is possible that there could be other remnant populations of the summit rat on other high mountains in northern Borneo (Appendix 2.7). Any of these scenarios is consistent with the summit rat populations on Mt. Tambuyukon and Mt. Kinabalu sharing a large, continuous ancestral population, as indicated by the estimation of an effective population size of the ancestral population probably at around twice as large as the one estimated for Kinabalu. We observed lower genetic diversity in the summit rats from Mt. Tambuyukon than those from Mt. Kinabalu, likely a consequence of drift after the retreat of the summit rat to higher elevations in the present interglacial. A very divergent mitochondrial haplotype (sample BOR 161), which was 131 substitutions apart from the closest haplotype in the network, despite all others haplotypes being only 1 to 18 bases apart, was identified in the Mt. Tambuyukon population. This is not a misidentified animal, as it was not divergent or differentiated at any nuclear locus (Figure 2.1 D). This divergent mitochondrial haplotype could be ancestral polymorphism, which is compatible with the demographic history supported by the ABC analyses in which there was a large ancestral population size followed by fragmentation and reduction in population size. Alternatively, this could be a case of mitochondrial introgression from another distant, unknown population or from its lowland sister species, *Rattus tiomanicus* (Aplin *et al.,* 2011).

Predicted responses to ongoing climate change

Elevational shifts induced by historic climate change have been previously observed in animal communities (Parmesan, 2006). For instance, half of the small mammal community in Yosemite, US, have shifted their ranges an average of 500 m uphill in one century (Moritz *et al.,* 2008). The effects of climate change do not manifest equally across the globe and are difficult to predict for a specific location because they depend not only on the concentration of gases with greenhouse effect but also on complex atmosphere-ocean global circulation models (IPCC, 2013). A re-survey of the moth community on Mt. Kinabalu showed that climate warming caused it to shift uphill by 67 m in just 42 years (Chen *et al.,* 2009), and a review of recent and historical distribution of birds in Mt. Kinabalu points to a similar effect in their community (Harris *et al.,* 2012). As climate change progresses, populations on some mountains may find themselves trapped at the top of a mountain and unable to disperse to another one (Sauer *et al.,* 2011). Scenarios with CO_2 concentrations of 690 ppm (twice the concentration at the end of the last century) predict an increase in temperature of about +2.2 ºC, and a shift uphill of the tropical cloud forest by 492 m (Still *et al.,* 1999). These CO_2 concentrations best match mild IPCC scenarios RCP6.0 for 2100 (IPCC, 2013), and the change in temperature predicted is slightly below the +2.3 – +3.1 ºC range estimated by different models for year 2085 (Nogués-Bravo *et al.,* 2007). Assuming an increase of 2.3 to 3.1 ºC on Mt. Kinabalu and a lapse rate of 5 ºC/Km (Sekercioglu *et al.,* 2008), by the end of this century the mountain communities in northern Borneo will have shifted up the mountains 460-620 m in elevation, reducing the habitat available to species dependent on montane forests, such as the summit rat.

The consequent reduction in the distribution area for summit rats is likely to reduce population sizes and increase drift, eroding levels of genetic diversity both on Mt. Kinabalu and Mt. Tambuyukon. This genetic erosion can be particularly severe for genetically isolated populations such as species with sky-island distributions (Bálint *et al.,* 2011). Assuming the distribution of the summit rat tightly follows the elevation shift in the thermal isoline derived from climate warming, the population isolated on Mt Tambuyukon and its unique genetic variation will likely disappear by the end of the current century. However, the deleterious effects of global warming will likely be milder on Mt. Kinabalu, where the height of the mountain provides room for an

elevational shift to a higher elevation. This prediction should be applicable to all species with limited dispersal capabilities that are tightly associated with this habitat.

Acknowledgements

We thank all the staff that participated in the fieldwork. Ch'ien C. Lee kindly provided beautiful summit rat pictures. Anna Cornellas helped with labwork. Vicente García Navas, Carles Vilà, Santiago Montero, Giovanni Forcina, Juanma Peralta, Inés Sánchez Donoso and other members of CONSEVOL (www.consevol.org), Lillian Parker, Nancy McInerney, Rob Fleischer and other members of the CCG at SCBI (Smithsonian National Zoo), and Kristofer Helgen kindly provided insight. We further thank the Malaysian institutions that allowed us to do fieldwork and export the samples: the Sabah Biodiversity Centre for issuing a research and export permit, Sabah Parks for research and collection permits, as well as support and cooperation during our time in Malaysia, and the Sabah Wildlife Department and the Economic Planning Unit for research and export permits. Logistical support was provided by Laboratorio de Ecología Molecular, Estación Biológica de Doñana, CSIC (LEM-EBD). Digital elevation models for maps was obtained from the U.S. Geological Survey. The Spanish Ministry of Science and Innovation grants CGL2010-21524 and CGL2014-58793-P supported this work. MCS is supported by the Spanish Ministry of Science and Innovation Predoctoral Fellowship BES-2011-049186 and part of his fieldwork was also funded by EEBB-I-12-05317.

Literature cited

Allendorf, F.W. & Luikart, G. (2007) Conservation and the Genetics of Populations. Blackwell Publishing Ltd.

Aplin, K.P., Suzuki, H., Chinen, A.A., *et al.* (2011) Multiple Geographic Origins of Commensalism and Complex Dispersal History of Black Rats. *PLoS ONE*, **6**, e26357.

Aplin, K. (2016) *Rattus baluensis. The IUCN Red List of Threatened Species 2016: e.T19323A22443731. Downloaded on 21 November 2016.*

Bálint, M., Domisch, S., Engelhardt, C.H.M., *et al.* (2011) Cryptic biodiversity loss linked to global climate change. *Nature Climate Change*, **1**, 313–318.

Barkman, T.J. & Simpson, B.B. (2001) Origin of High-Elevation *Dendrochilum* Species (Orchidaceae) Endemic to Mount Kinabalu, Sabah, Malaysia. *Systematic Botany*, **26**, 658–669.

Beaman, J.H. (2005) Mount Kinabalu: Hotspot of plant diversity in Borneo. *Biologiske Skrifter*, **55**, 103–127.

Beaumont, M.A., Zhang, W. & Balding, D.J. (2002) Approximate Bayesian computation in population genetics. *Genetics*, **162**, 2025–2035.

Bennett, K.D. & Provan, J. (2008) What do we mean by "refugia"? *Quaternary Science Reviews*, **27**, 2449–2455.

Brandariz-Fontes, C., Camacho-Sanchez, M., Vilà, C., *et al.* (2015) Effect of the enzyme and PCR conditions on the quality of high-throughput DNA sequencing results. *Scientific Reports*, **5**, 8056.

Camacho-Sanchez, M., Burraco, P., Gomez-Mestre, I. & Leonard, J.A. (2013) Preservation of RNA and DNA from mammal samples under field conditions. *Molecular Ecology Resources*, **13**, 663–673.

Cannon, C.H., Morley, R.J. & Bush, A.B.G. (2009) The current refugial rainforests of Sundaland are unrepresentative of their biogeographic past and highly vulnerable to disturbance. *Proceedings of the National Academy of Sciences*, **106**, 11188–11193.

Castresana, J. (2000) Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Molecular Biology and Evolution*, **17**, 540–552.

Chen, I.-C., Shiu, H.-J., Benedick, S., *et al.* (2009) Elevation increases in moth assemblages over 42 years on a tropical mountain. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 1479–1483.

Chin, L., Moran, J.A. & Clarke, C. (2010). Trap geometry in three giant montane pitcher plant species from Borneo is a function of tree shrew body size. *New Phytologist*, **186**, 461-470.

Clarke, C.M., Bauer, U., Lee, C.C., *et al.* (2009) Tree shrew lavatories: a novel nitrogen sequestration strategy in a tropical pitcher plant. *Biology Letters*, **5**, 632–5.

Clement, M., Posada, D. & Crandall, K.A. (2000) TCS: A computer program to estimate gene genealogies. *Molecular Ecology*, **9**, 1657–1659.

Darriba, D., Taboada, G.L., Doallo, R. & Posada, D. (2012) jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods*, **9**, 772–772.

Drummond, A.J., Suchard, M.A., Xie, D. & Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*, **29**, 1969-1973.

Earl, D.A. & vonHoldt, B.M. (2012) STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, **4**, 359–361.

Evanno, G., Regnaut, S. & Goudet, J. (2005) Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Molecular Ecology*, **14**, 2611–2620.

Frankham, R. (1995). Effective population size/adult population size ratios in wildlife: a review. *Genetical Research*, **66**, 95-107.

Greenwood, M., Clarke, C., Ch'ien, C.L., Gunsalam, A. & Clarke, R.H. (2011). A unique resource mutualism between the giant Bornean pitcher plant, *Nepenthes rajah*, and members of a small mammal community. *PLoS One*, **6**, e21114.

Harris, J.B.C., Sheldon, F.H., Boyce, A.J., *et al.* (2012) Using diverse data sources to detect elevational range changes of birds on Mount Kinabalu, Malaysian Borneo. *The Raffles Bulletin of Zoology*, **25**, 197–247.

Hawkins, M. T. R. 2015. Biogeography and Phylogeography of Mammals of Southeast Asia: A Comparative Analysis Utilizing Macro and Microevolution. Doctoral thesis. George Mason University.

He, K., Hu, N.-Q., Chen, X., Li, J.-T. & Jiang, X.-L. (2016) Interglacial refugia preserved high genetic diversity of the Chinese mole shrew in the mountains of southwest China. *Heredity*, **116**, 23–32.

He, K. & Jiang, X. (2014) Sky islands of southwest China. I: An overview of phylogeographic patterns. *Chinese Science Bulletin*, **59**, 585–597.

Hewitt, G. (1996) Some genetic consequences of ice ages, and their role in divergence and speciation. *Biological Journal of the Linnean Society*, **58**, 247–276.

Hewitt, G. (1999) Post-glacial re-colonization of European biota. *Biological Journal of the Linnean Society*, **68**, 87–112.

Hewitt G (2000) The genetic legacy of the Quaternary ice ages. *Nature*, **405**, 907–913.

Hope, G.S. (2007) Paleoecology and Paleoenvironments of Papua. In: *The Ecology of Papua: Part One* (eds Marshall AJ, Beehler BM), pp. 255–266. Tuttle Publishing, Singapore.

Igea, J., Juste, J. & Castresana, J. (2010) Novel intron markers to study the phylogeny of closely related mammalian species. *BMC Evolutionary Biology*, **10**, 369.

IPCC (2013) *IPCC, 2013: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Jones, K.E., Bielby, J., Cardillo, M., *et al.* (2009) PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals. *Ecology*, **90**, 2648–2648.

Katoh, K. & Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, **30**, 772–780.

Kebede, M., Ehrich, D., Taberlet, P., Nemomissa, S. & Brochmann, C. (2007) Phylogeography and conservation genetics of a giant lobelia (*Lobelia giberroa*) in Ethiopian and Tropical East African mountains. *Molecular Ecology*, **16**, 1233–1243.

Kimura, Y., Hawkins, M.T.R., Mcdonough, M.M., Jacobs, L.L. & Flynn, L.J. (2015) Corrected placement of *Mus* - *Rattus* fossil calibration forces precision in the molecular tree of rodents. *Scientific Reports*, **5**, 14444.

Kitayama, K. (1992). An altitudinal transect study of the vegetation on Mount Kinabalu, Borneo. *Vegetatio*, **102**, 149-171.

Kopelman, N.M., Mayzel, J., Jakobsson, M., Rosenberg, N.A. & Mayrose, I. (2015) Clumpak: A program for identifying clustering modes and packaging population structure inferences across K. *Molecular Ecology Resources*, 1179–1191.

Librado, P. & Rozas, J. (2009) DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, **25**, 1451–1452.

Lopes, J.S., Balding, D. & Beaumont, M.A. (2009) PopABC: a program to infer historical demographic parameters. *Bioinformatics*, **25**, 2747–9.

MacVean, C. & Schuster, J.C. (1981) Altitudinal Distribution of Passalid Beetles (Coleoptera, Passalidae) and Pleistocene Dispersal on the Volcanic Chain of Northern Central America. *Biotropica*, **13**, 29–38.

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**, 10–12.

Merckx, V.S.F.T., Hendriks, K.P., Beentjes, K.K., *et al.* (2015) Evolution of endemism on a young tropical mountain. *Nature*, **524**, 347–350.

Moritz, C., Patton, J.L., Conroy, C.J., *et al.* (2008) Impact of a century of climate change on small-mammal communities in Yosemite National Park, USA. *Science*, **322**, 261–264.

Musser, G.G. (1986) Sundaic *Rattus*: definitions of *Rattus baluensis* and *Rattus korinchi*. *American Museum Novitates*, **2862**, 1–24.

Nogués-Bravo, D., Araújo, M.B., Errea, M.P. & Martínez-Rica, J.P. (2007) Exposure of global mountain systems to climate warming during the 21st Century. *Global Environmental Change*, **17**, 420–428.

Nor, S.M. (2001) Elevational diversity patterns of small mammals on Mount Kinabalu, Sabah, Malaysia. *Global Ecology and Biogeography*, **10**, 41–62.

Pacifici. M., Santini. L., Di Marco, M., *et al.* (2013) Generation length for mammals. *Nature Conservation*, **5**, 87–94.

Parmesan, C. (2006) Ecological and Evolutionary Responses to Recent Climate Change. *Annual Review of Ecology, Evolution, and Systematics*, **37**, 637–669.

Peakall, R. & Smouse, P.E. (2012) GenALEx 6.5: Genetic analysis in Excel. Population genetic software for teaching and research. *Bioinformatics*, **28**, 2537–2539.

Petit, R.J., Aguinagalde, I., de Beaulieu, J.-L., *et al.* (2003) Glacial refugia: hotspots but not melting pots of genetic diversity. *Science*, **300**, 1563–1565.

Phillipps, Q. & Phillipps, K. (2016) *Phillips' field guide to the mammals of Borneo and their ecology*. Natural History Publications (Borneo), Kota Kinabalu.

Pickett, E.J., Harrison, S.P., Hope, G., *et al.* (2004). Pollen-based reconstructions of biome distributions for Australia, Southeast Asia and the Pacific (SEAPAC region) at 0, 6000 and 18,000 14C yr BP. *Journal of Biogeography*, **31**, 1381-1444.

Porter, S.C. (2001) Snowline depression in the tropics during the last glaciation. *Quaternary Science Reviews*, **20**, 1067–1091.

Pritchard, J.K., Stephens, M. & Donnelly, P. (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–59.

QGIS Development Team (2015) QGIS Geographic Information System. *Open Source Geospatial Foundation Project*.

R Core team (2015) R Core Team. *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.*, **55**, 275–286.

Raes, N., Roos, M.C., Slik, J.W.F., Van Loon, E.E., Ter Steege, H. (2009) Botanical richness and endemicity patterns of Borneo derived from species distribution models. *Ecography*, **32**, 180–192.

Sauer, J., Domisch, S., Nowak, C. & Haase, P. (2011) Low mountain ranges: summit traps for montane freshwater species under climate change. *Biodiversity and Conservation*, **20**, 3133–3146.

Sekercioglu, C.H., Schneider, S.H., Fay, J.P. & Loarie, S.R. (2008) Climate change, elevational range shifts, and bird extinctions. *Conservation Biology*, **22**, 140–150.

Sikes, R.S., Gannon, W.L. & the Animal Care and Use Committee of the American Society of Mammalogists (2011) Guidelines of the American Society of Mammalogists for the use of wild mammals in research. *Journal of Mammalogy*, **92**, 235–253.

Smythies, B.E. (1964) The birds of Mt Kinabalu and their zoogeographical relationships. *Proceedings of the Royal Society of London B: Biological Sciences,* **161**, 75-80.

Stauffer, P.H. (1968) Glaciation of Mount Kinabalu. *Geological Survey of Malaysia Bulletin*, **1**, 63.

Stewart, J. R., Lister, A. M., Barnes, I., & Dalén, L. (2010). Refugia revisited: individualistic responses of species in space and time. *Proceedings of the Royal Society of London B: Biological Sciences*, **277**, 661-671.

Still, C.J., Foster, P.N. & Schneider, S.H. (1999) Simulating the effects of climate change on tropical montane cloud forests. *Nature*, **398**, 15–17.

Struebig, M.J., Wilting, A., Gaveau, D.L.A., *et al.* (2015). Targeted Conservation to Safeguard a Biodiversity Hotspot from Climate and Land-Cover Change. *Current Biology*, **25**, 372-378.

van der Ent, A., Repin, R., Sugau, J. & Meng Wong, K. (2015) Plant diversity and ecology of ultramafic outcrops in Sabah (Malaysia). *Australian Journal of Botany*, **63**, 204–215.

Wells, K., Lakim, M.B., Schulz, S. & Ayasse, M. (2011a). Pitchers of *Nepenthes rajah* collect faecal droppings from both diurnal and nocturnal small mammals and emit fruity odour. *Journal of Tropical Ecology*, **27**, 347-353.

Wells, K., Lakim, M.B. & Beaucournu, J.C. (2011b). Host specificity and niche partitioning in flea–small mammal networks in Bornean rainforests. *Medical and Veterinary Entomology*, **25**, 311–319.

Appendix 2.1. Amplicon library preparation of intron markers

We amplified a panel of introns for the 47 field samples. These introns were proposed by Igea *et al.* (2010) to study phylogenies of closely related non-rodent mammal species, although their variation at the intra-species level was unknown. We designed primer pairs for 30 of these introns based on alignments of the *Mus musculus* (mm8) and *Rattus norvegicus* (Rnor_5.0) genomes to prepare Roche 454 GS Junior amplicon libraries following the universal tailed amplicon sequencing design (Roche 2011). We first did a test run and selected a panel of 27 primer pairs that amplified target introns in the summit rat (Table 2.2). The 27 selected loci were amplified in a single polymerase chain reaction (PCR) using the Multiplex PCR Kit (Qiagen). PCR reactions consisted of 1x master mix and ~75 ng of DNA in a final volume of 25 µl. The primer concentrations varied from 0.11 to 0.35 µM (Table 2.2) according to a previous test run on the Roche 454 GS Junior to determine relative amplification of each locus. The PCR program started with an initial denaturation at 95 ºC for 15 min, followed by 22 cycles of a denaturation at 95ºC for 30 s, annealing at 59 ºC for 90 s and extension at 72 ºC for 2 min. These conditions, with reduced number of cycles, long extension time and no final extension were selected to reduce sequence artifacts (Brandariz-Fontes *et al.,* 2015). The PCR products were cleaned using SPRI beads (Rohland & Reich, 2012) and checked on agarose gels. The diluted templates (from 1:2 to 1:20, depending on the sample) were used in a second PCR to ligate an adaptor with an individual index. Each reaction consisted of 1.67 µM of indexing oligos, 1x Phusion Master Mix (New England Biolabs) and 2 µl of diluted PCR product in 12 µl total volume. PCR conditions were as follows: initial denaturation at 98ºC for 30 s, followed by 25 cycles of 98ºC for 10 s, 56ºC for 20 s and 72ºC for 45 sec. PCR products were quantified on an agarose gel with a reference standard in Quantity-One software (Bio-Rad Laboratories). Three samples were replicated. The libraries were pooled at equal concentrations and double-cleaned with SPRI beads. An emulsion PCR was performed with the GS Junior Titanium emPCR Kit Lib-A (Roche) and sequenced in a Roche 454 GS Junior. Of the 27 introns, three of them (*rgd735029*-5, *mmp9*-2 and *mycbpap*-11) had very low number of reads and were re-sequenced for all 47 individuals, along with eight samples that were also re-sequenced for all introns due to insufficient coverage.

Genotyping of nuclear markers

Data from the Roche 454 GS Junior runs were imported into Geneious R6 (http://www.geneious.com, Kearse *et al.,* 2012) and reads shorter than 250 bp were removed. Reads were demultiplexed by barcode and primers were trimmed. For each sample reads were mapped to a reference sequence from *Rattus norvegicus* (Rnor_5.0). Genotype calling was done manually, considering only those loci with a minimum of 10x coverage. To ensure reliability in calling heterozygotes, only alleles present in both forward and reverse reads and that had a minimum of 4x coverage were considered.

Mitochondrial DNA sequencing

We sequenced complete mitochondrial genomes for all 49 samples (Table 2.1). Before library preparation, the DNA samples were sonicated in a Bioruptor UCD-200 (Diagenode) in 100 µl at 20 ng/µl in 1.5 ml Eppendorf tubes with 3-6 cycles of 30 s on/off with frequency set to high to target mean fragment sizes of 300-400 bp.

A subset of 25 samples was indexed and directly sequenced with no enrichment. In this case, we used half reactions of NEBNext DNA Library Prep Master Mix Set (New England Biolabs) following manufacturer's instruction, except that the cleaning steps were performed with QIAquick Spin Columns (Qiagen) instead of beads. We amplified the libraries in 25 µl with 0.2 mM each dNTP, 2 mM of $MgSO_4$, 0.5 µM of PE 1.0 primer (AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC*T), 0.5 µM of PE 2.0 primer (GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T), 0.1 µM of an indexing oligo p7 (CAAGCAGAAGACGGCATACGAGAT-index-GTGACTGGAGTTCAGACGTGTGCTCTTCCG), 1x Platinum buffer, 1.25 U Platinum HiFi Polymerase (Invitrogen, Life Technologies) and 125 ng of template. The PCR program started with an initial denaturation at 94 ºC for 5 min, 10 cycles of 30 s at 94ºC, 30 s at 60 ºC and 40 s at 68 ºC, and a final elongation step of 7 min at 68 ºC. The products were cleaned with 1x SPRI beads and quantified by qPCR with Library Quantification Kit - Illumina/Universal (Kapa Biosystems) on a Stratagene Mx3005p (Agilent Technologies). We pooled the libraries at equimolar ratios and sequenced them on an Illumina MiSeq using 250 bp Paired-End chemistry at the Danish National High-throughput Sequencing Centre (University of Copenhagen, Denmark).

Sequences derived from liver samples in this run had sufficient reads to reconstruct complete mitogenomes, but those derived from ear tissue did not. For this subset of samples, we utilized a hybridization protocol to enrich the libraries for mitochondrial genomes. These libraries were prepared as in Meyer and Kircher (2010) with the Kapa Illumina Library Preparation Kit (Kapa Biosystems) with some modifications (methods in Chapter 4). Briefly, we prepared double-indexed shotgun libraries which we enriched for mitochondrial DNA and sequenced on an Illumina HiSeq 2000 with 100 bp paired-end reads at Macrogen. Three samples were replicated using the different library preparation protocols to ensure repeatability.

We batch processed the sequences following a custom pipeline (below). First, we trimmed Illumina adaptors and low quality regions with Cutadapt 1.8.3 (-a AGATCGGAAGAGC -e 0.16 -O 1 -q 25) (Martin 2011). We generated a reference mitogenome for the summit rat by mapping the reads from sample BOR373 to a complete mitogenome of *Rattus norvegicus* (GenBank: AJ428514). We selected BOR373 because it originated from a non-enriched library, to avoid potential biases associated with the enrichment. The mapping was done in Geneious 8.1.5 using medium-low sensitivity and three iterations. We generated a consensus sequence in Geneious and then indexed the summit rat reference using BWA 0.7.12-r1039. This was used as a reference to map R1 and R2 reads of all the other samples using the BWA-MEM algorithm (Li, 2013). We removed PCR duplicates from the *sam* mapping files using samtools 1.3 (Li *et al.,* 2009). First, *sam* mapping files were converted to *bam* and only the mapped reads were kept. Then, the *sam* files were sorted and the duplicates were removed. Using samtools we merged the mapping files from individuals that had been sequenced using different strategies to increase their coverage. This was not done for the three replicate samples. The resulting mapping *sam* files with duplicates removed were imported into Geneious 8.1.5 and the coverage and quality of the mapping were checked visually. From them we generated consensus sequences in Geneious using a 75% threshold and a minimum of 2x coverage. Only mitogenomes with more than 99.5% of their sequence reconstructed were considered for downstream analysis, yielding a total of 32 mitogenomes out of the 49 samples (Table 2.1) that ranged from 16,308 to 16,313 bp.

Mitochondrial genome assembly pipeline

Bash script for batch assembly of complete mitochondrial genomes from Illumina reads. All programs (*cutadapt*, *BWA* and *samtools*) should be added to the $PATH.

After each step the files generated might have long names which can be simplified by batch renaming the files, replacing a long *pattern* with a simpler one:

```
for f in pattern;
do
a="$(echo $f | sed s/pattern/pattern_to_replace_with/)"
mv "$f" "$a"
done
```

1. Trim adaptors with cutadapt.

```
for file in *.fastq
do
cutadapt -a AGATCGGAAGAGC -e 0.16 -O 1 -q 25 "$file" > "$file"_cutadapt.fastq
done
```

Cutadapt will trim adaptors from the 3' end for all *fastq* reads with a maximum sequence dissimilarity of 16 % respect to illumina sequence adaptor –a, including partial matches. The –q 25 argument trims low-quality bases from the 3' end.

2. Create a good quality reference. *BWA* requires a close reference to map the reads. If there is no close reference available you might opt for an iterative mapping algorithm instead.

3. Index the reference with *BWA*.

```
bwa index reference.fasta
```

4. Batch mapping the reads to the indexed reference with *BWA*.

```
fastq1=($(find *2_cutadapt.fastq))
fastq2=($(find *1_cutadapt.fastq))
long=($(echo ${#fastq1[@]}))
for i in `seq 0 $((long-1))`
do
bwa mem reference.fasta ${fastq1[i]} ${fastq2[i]} > ${fastq1[i]}.sam
done
```

BWA maps R1 reads (*1_cutadapt.fastq)* and R2 reads (*2_cutadapt.fastq*) to an indexed reference. It can be easily modified to be used with single-end reads.

5. Remove PCR duplicates on the mapped files with *samtools*.

5.1 Convert mapping files from SAM to BAM and keep only mapped reads.

```
for file in *.sam
do
samtools view -Shu -F 4 "$file" > "$file".bam
done
```

5.2 Sort bam files:

```
for file in *.bam
do
samtools sort "$file" -o "$file"_sorted
done
```

5.4 Remove PCR duplicates:

```
for file in *sorted
do
samtools rmdup -S "$file" "$file"_rmdup.bam
done
```

Literature cited in Appendix 2.1:

Brandariz-Fontes, C., Camacho-Sanchez, M., Vilà, C., *et al.* (2015) Effect of the enzyme and PCR conditions on the quality of high-throughput DNA sequencing results. *Scientific Reports*, **5**, 8056.

Igea, J., Juste, J. & Castresana, J. (2010) Novel intron markers to study the phylogeny of closely related mammalian species. *BMC Evolutionary Biology*, **10**, 369.

Kearse, M., Moir, R., Wilson, A., *et al.* (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, **28**, 1647–1649.

Li, H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *Preprint at arXiv:1303.3997v2 [q-bio.GN]*.

Li, H., Handsaker, B., Wysoker, A., *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

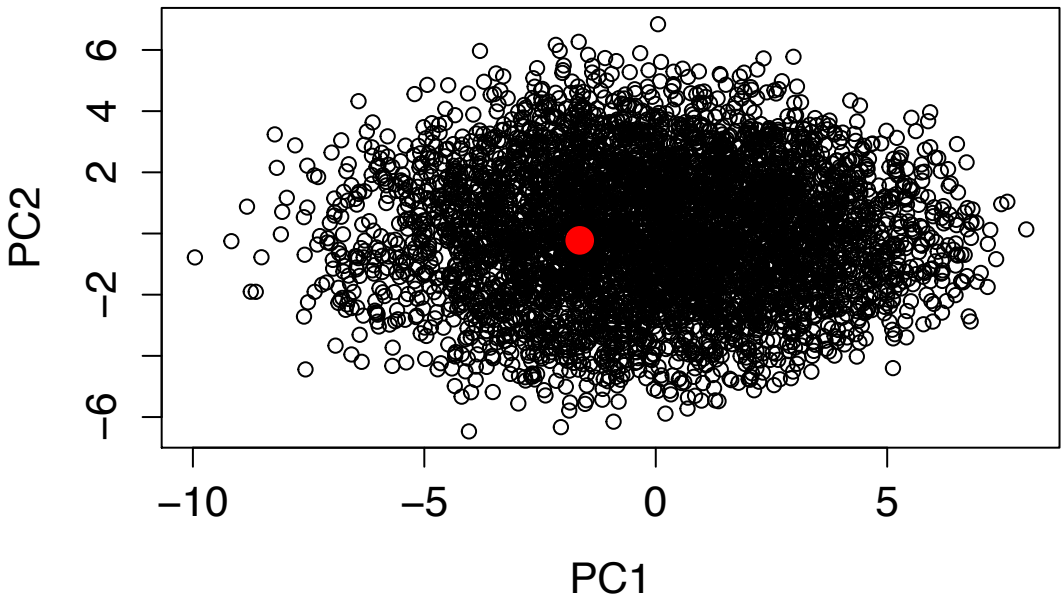
Meyer, M. & Kircher, M. (2010) Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols*, **2010**, pdb.prot5448.

Roche (2011) 454 Sequencing System Guidelines for Amplicon Experimental Design.

Rohland, N. & Reich, D. (2012) Cost-effective , high-throughput DNA sequencing. *Genome Research*, **22**, 939–946.

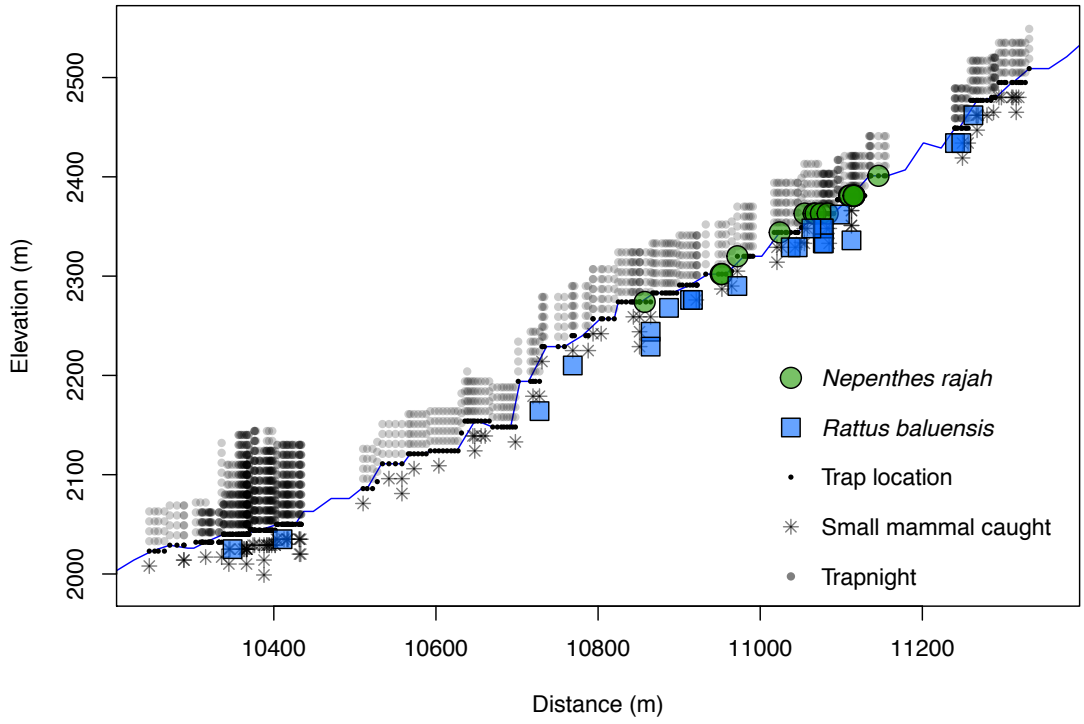
Appendix 2.2. PCA of the rejection step by PopABC

PCA of the rejection file produced by PopABC. The red dot corresponds to the study data and the open circles to the top 10 % of the simulated data.



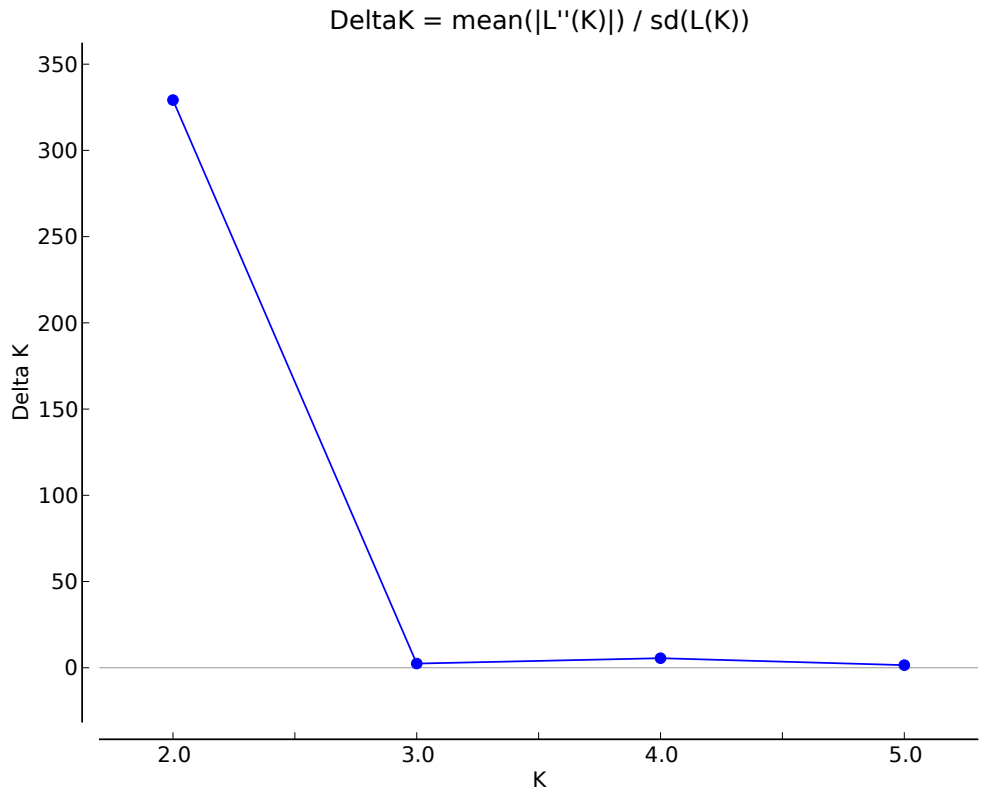
Appendix 2.3. *Nepenthes rajah* and summit rat catches on Tambuyukon

Location of *Nepenthes rajah* and summit rat catches on Tambuyukon. Trapping effort, summit rat occurrence and *Nepenthes rajah* observations on Mt. Tambuyukon. We recorded all *Nepenthes rajah* pitcher plants along the trap transects above 2,000 m on Mt. Tambuyukon. The areas where *Nepenthes rajah* was more abundant correlated with the areas where we trapped more summit rats.



Appendix 2.4. Evanno Delta K

Delta K calculated by Structure HARVESTER following the Evanno method (Evanno *et al.* 2005).

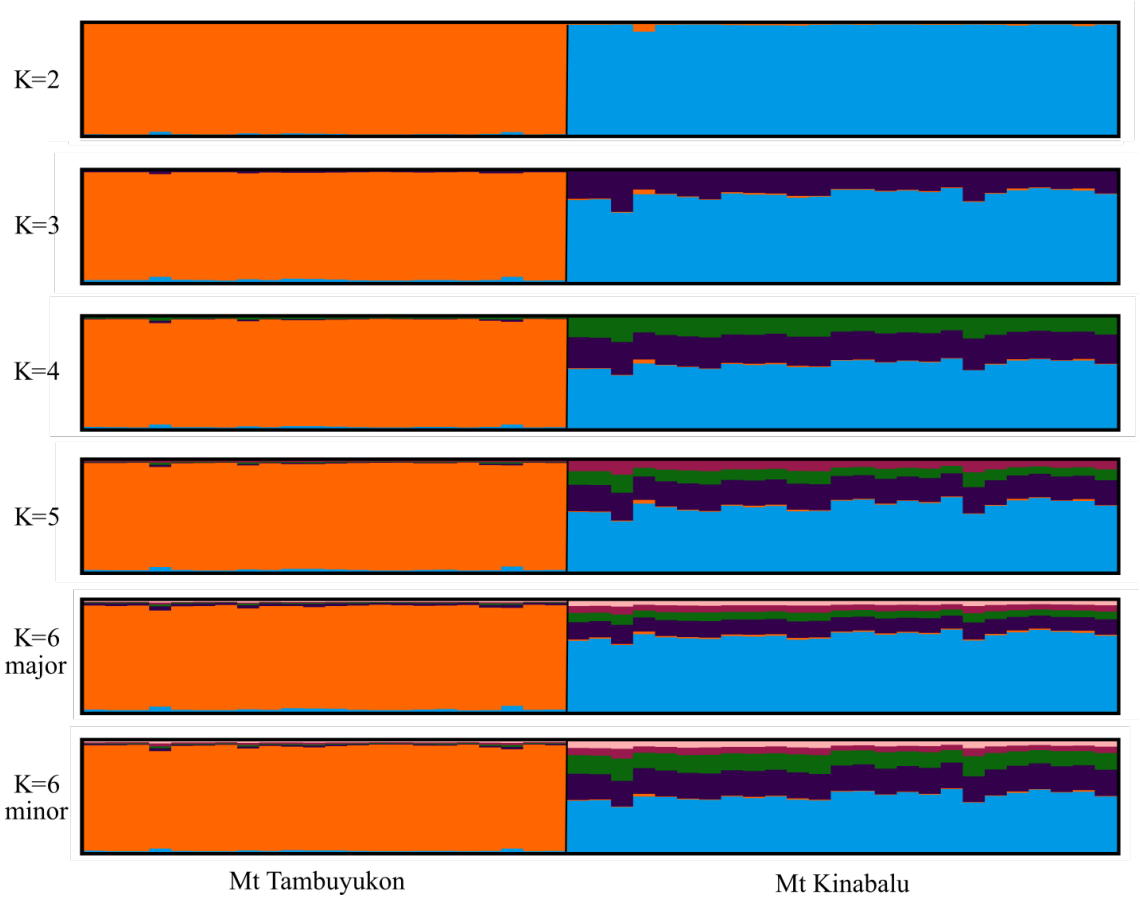


Literature cited in Appendix 2.4:

Evanno, G., Regnaut, S. & Goudet, J. (2005) Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Molecular Ecology*, **14**, 2611–2620.

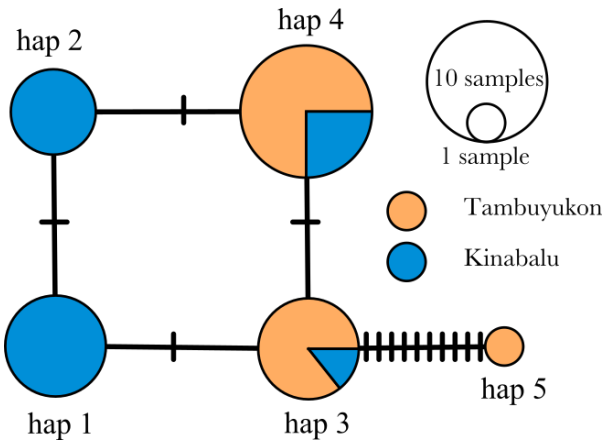
Appendix 2.5. K2 to K6 from STRUCTURE

STRUCTURE results using nuclear intron sequences from K2 to K6 for the four independent runs and later clustered with CLUMPAK.

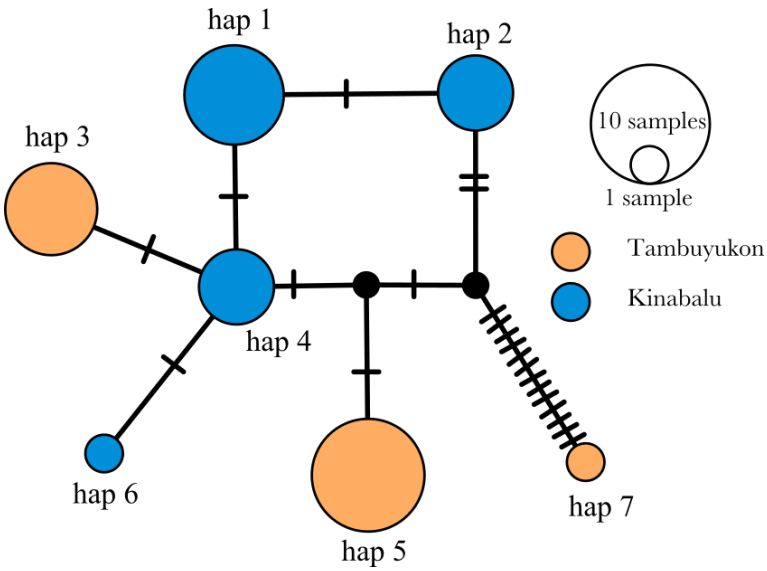


Appendix 2.6. *cytb* and control region haplotype networks

TCS haplotype networks for mitochondrial *cytochrome b* (top) and control region (bottom) haplotypes from summit rats of Mt. Tambuyukon (orange) and Mt. Kinabalu (blue).



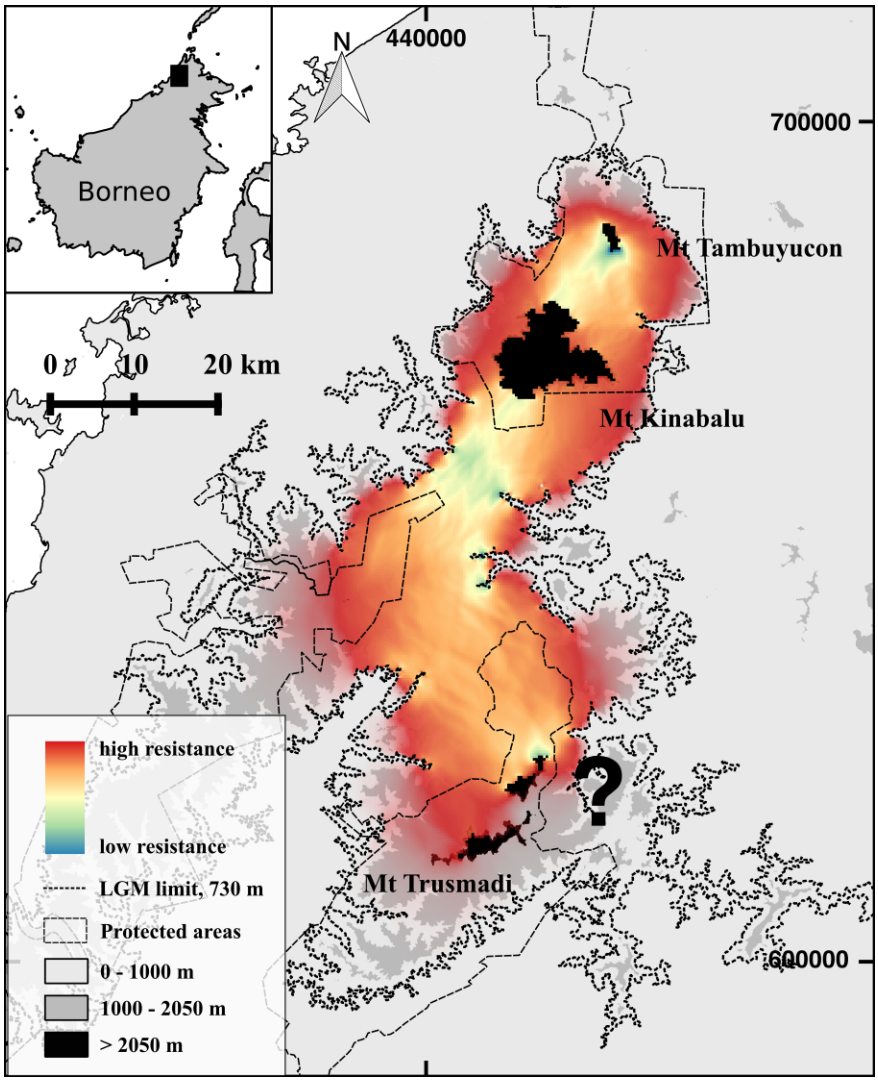
TCS haplotype network for *cytochrome b*.



TCS haplotype network for the control region.

Appendix 2.7. Connectivity reconstructions

Possible connectivity between summit rat populations in the late Pleistocene. During the Last Glacial Maximum, 21 Kya, the lower distribution limit for the summit rat is could have shifted from its today's 2,040 m down to 720 m, dotted line; see main text for details. Today's climatic conditions are unusual, as glaciation conditions prevailed for 90 % of the Pleistocene (Stewart *et al.* 2010; Woodruff 2010) implying a larger distribution for the summit rat than today, as indicated by a larger ancestral population size (see main text for details). We predicted the connectivity between the focal nodes of the current distribution of summit rats on Mt. Kinabalu and Mt. Tambuyukon, and considered other mountains over 2,000 m in Sabah as well. We used Circuitscape 4.0.5, that borrows algorithms from the electronic circuit theory (Mcrae *et al.* 2008) to estimate the most likely pathways of gene flow between the mountains through glaciations. As a resistance matrix we used SRTM digital elevation models from the USGS (http://earthexplorer.usgs.gov/). In Circuitscape, we did pairwise iterations across all pairs of focal nodes treating elevations in the raster data as conductances instead of resistances, and the conductance of elevations below 720 m was set to 0.



Map showing most likely pathways of gene flow in summit rats between high altitude areas of Sabah through Pleistocene glaciations as estimated with Circuitscape.

Literature cited in Appendix 2.7:

Mcrae, B.H., Dickson, B.G., Keitt, T.H. & Shah, V.B. (2008) Using Circuit Theory to Model Connectivity in Ecology, Evolution, and Conservation. *Ecology*, **89**, 2712–2724.

Stewart, J.R., Lister, A.M., Barnes, I. & Dalén, L. (2010) Refugia revisited: individualistic responses of species in space and time. *Proceedings of The Royal Society B*, **277**, 661–71.

Woodruff, D.S. (2010) Biogeography and conservation in Southeast Asia: how 2.7 million years of repeated environmental fluctuations affect today's patterns and the future of the remaining refugial-phase biodiversity. *Biodiversity and Conservation*, **19**, 919–941.

Chapter 3 Rapid external morphological divergence after mountain colonization in a Sunda rat

Miguel Camacho Sánchez[1*], Kristofer M. Helgen[2] and Jennifer A. Leonard[1]

[1]Conservation and Evolutionary Genetics Group, Estación Biológica de Doñana (EBD-CSIC), Avda Américo Vespucio 26, 41092, Sevilla, Spain

[2] *current address: School of Biological Sciences, University of Adelaide, Adelaide, South Australia 5005, Australia.

[3]Division of Mammals, National Museum of Natural History, Smithsonian Institution, P.O. Box 37012, Washington DC 20013-7012, USA.

Abstract

Selective forces driving morphological evolution in tropical montane environments may act similarly on the various, distantly related small mammals endemic to these habitats. The apparent convergent adaptation to this habitat has confounded several previous taxonomic studies, or simply left some relationships as undetermined. Here we use whole mitochondrial genomes from the three high altitude *Rattus* species that are endemic to Sundaland (*Rattus korinchi* and *R. hoogerwerfi* from Sumatra; and *R. baluensis* from Borneo), and several related species in order to determine their relationships, and date divergences with sister taxa. The two montane species from Sumatra were on long branches within the Asian *Rattus*. The high altitude species from Borneo shared external morphological traits with the other mountain species, but its origin is very recent (< 390 Kya) and nested within the diversity of *R. tiomanicus*, a widespread lowland species. These data suggest selective forces and/or "island effect" associated with mountain habitats drive convergent evolution on an evolutionary short timescale.

Introduction

Mountains offer ecological clines along which selection pressures can have different strengths or even opposite directions. In endothermic vertebrates, such as mammals, cold and hypoxic conditions in high mountain environments can drive genetic and consequent physiological adaptive changes (Cheviron and Brumfield 2012). This has been mainly shown for pathways related to oxygen affinity and thermogenesis (Storz 2007; Storz et al. 2009; Beall et al. 2010; Cheviron et al. 2011; Cheviron et al. 2013), or even to protection from pulmonary injury (Gorkhali et al. 2016). Other adaptations involving external morphology, such as longer and denser fur have also been observed (Wasserman and Nash 1979).

The biogeographical region of Sundaland, in Southeast Asia, hosts four endemic species of *Rattus* (Musser and Newcomb 1983; Musser and Carleton 2005), three of which are endemic to high altitude: *Rattus baluensis* (Thomas, 1894), known only from Sabah, northern Borneo, and *R. korinchi* (Robinson & Kloss, 1916) and *R. hoogerwerfi* (Chasen, 1939), known only from a few records above 2,000 m on Sumatra (Robinson and Kloss 1918,1919; Miller 1942; Musser 1986; Chapter 2) (Figure 3.1). These three species inhabit similar mountain habitats, which seems to be driving convergence in external morphological characters as an adaptation to colder temperatures; mainly long dark fur with woolly underfur (Figure 3.2; Musser 1986). The evolutionary relationships between these Sunda rats are not clear. Their apparent convergence has misled taxonomists, and until 1986 *R. korinchi* was considered a subspecies of *Rattus baluensis* (Musser 1986). Chasen (1939) pointed to a potential affinity of *R. hoogerwerfi* to *R. korinchi*, although in a later examination of both species Musser (1986) declared they shared as many characters as those that separated them, thus they should be seen as singular lineages out of the *Rattus rattus* group, and that perhaps their relatives "*should be looked for in places off of the Sunda Shelf"* (Musser 1986:21). However, the evolutionary relationships at the molecular level between *R. hoogerwerfi* and *R. korinchi*, or to other *Rattus*, has never been assessed. In the most recent revision of *Rattus* they are considered among the species with unresolved phylogenetic affinities (Musser and Carleton 2005). Morphological traits and molecular data associate *R. baluensis* to the only lowland *Rattus* endemic to Sunda, *R. tiomanicus* (Miller, 1900), from which it seems to have originated (Musser and Newcomb 1983; Aplin et al 2011).

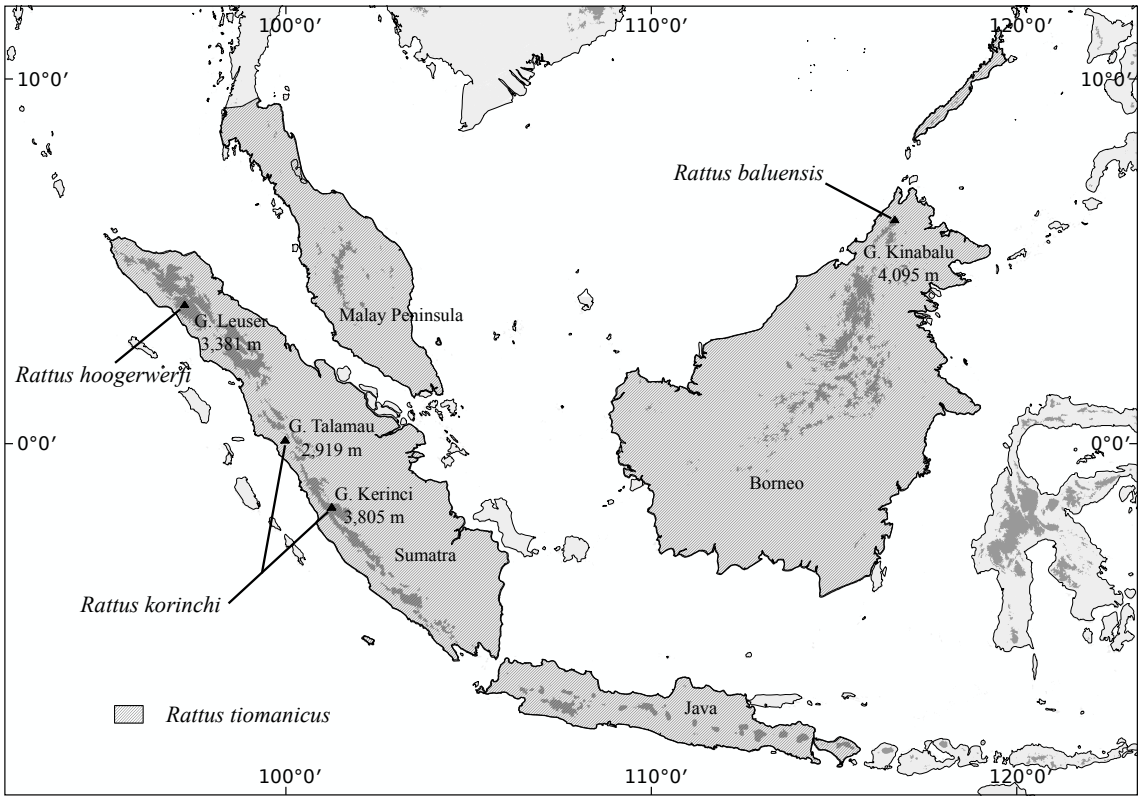


Figure 3-1. Distribution of the four *Rattus* species endemic to Sundaland. Dark grey corresponds to elevations above 1,000 m. Data for *R. tiomanicus* was downloaded from the IUCN (2016).

This study aimed to provide a phylogenetic framework to discuss convergence in external morphology associated to mountain *Rattus* in Sundaland. We constructed a dated mitogenome phylogeny from the three montane species of Sunda *Ratttus*, in which we also added the lowland *R. tiomanicus* and other *Rattus* species. These results are placed in the context of the evolution of other highland small mammals in the region.

Figure 3-2. Dorsal view of the skins of the lowland *R. tiomanicus*, *R. blangorum*, and the three montane endemics, *R. baluensis*, *R. hoogerwerfi*, and *R. korinchi*, with a (2.2x) detail of the woolly underfur of *R. korinchi*.

Methods

Study system

We include all endemic Sunda *Rattus* species recognized in Musser and Newcomb (1983), except for *R. annandalei*, which has been moved to *Sundamys annandalei* (Chapter 4). At present, other lowland generalists in Sunda, such as *R. argentiventer*, *R. rattus*, *R. norvegicus*, *R. tanezumi* or *R. exulans* are considered invasive. The *R. tiomanicus* complex includes several insular forms (*R. burrus, R. simalurensis, R. adustus, R. palmarum, R. mindoronensis* and *R. lugens*) which are recognized as different species (Musser and Carleton 2005), although their molecular affinity has not been addressed. *Rattus blangorum* was originally described from only two specimens (ANSP 20348 and 20349) from the Aceh region, northern Sumatra (Miller 1942), but it was placed later in the *R. tiomanicus* complex (Musser and Califia 1982; Musser and Carleton 2005). The three montane species (*R. baluensis*, *R. hoogerwerfi*, and *R. korinchi*) inhabit similar mountain habitat (Musser 1986). *R. baluensis* is found above 2,000 m in northern Borneo, from mossy forest, mountain scrubland and up to subalpine vegetation in Kinabalu (Musser 1986; Chapter 2). Musser and Carleton (2005) report 1,524 m as its lowest elevation, probably from a misidentified museum specimen reported in Nor (2001). *Rattus korinchi* is only known from two specimens collected in montane or moss forest on Mt. Kerinci at 2,164 m (Robinson and Kloss 1918; the holotype No. 442/14, =BM 19.11.5.81, and an immature female with no specimen tha we have been able to locate in museums) and Mt. Talamau at 2,773 m (Robinson and Kloss 1919; No. 351,=RMNH 23151, and No. 352 for which we have not been able to locate the museum specimen; Musser 1986). The other Sumatran mountain endemic, *R. hoogerwerfi*, is known from 29 specimens collected on Mt. Leuser, northern Sumatra, from 2,133-2,835 m, also inhabiting a similar montane habitat as its Kinabalu counterpart, described as "*moss forest, with trees averaging only 5 to 40 feet in height, very hard, knotted and twisted. Everywhere the ground and the branches of the trees were covered with a deep carpet of moss and ferns*" (Miller 1942:118) and also "*mostly bare or covered with grass interspersed with patches of bushes and low trees*" (Miller 1942:108), and seven specimens reported in Sody (1941:300), also from the same area. The 2,900 ft (884 m) where the holotype was collected (Chasen 1939; also cited in Musser and Carleton 2005) is probably an error as this species is confined to higher elevations (Miller 1942). The high trapping success for *R. hoogerwerfi* (Miller 1942)

compared to other small mammals suggests this species is present at high densities in its habitat, probably in the range for those reported for its Bornean counterpart *R. baluensis* on Mt. Kinabalu (Nor 2001; Chapter 2).

Taxonomic and gene sampling

We sequenced mitogenomes of modern and historical individuals from the Sundaland montane species *R. korinchi* and *R. hoogerwerfi*, the lowland species *R. tiomanicus*, and other representative species of *Rattus*, all of which were obtained from museum collections and the field (Table 3.1). One historical sample from Sabah labeled *Lenothrix canus* (NH 2015) was reassigned to *R. tiomanicus* based on *cytochrome b* (*cyt b*) barcoding. The historical specimen NH 2147 labeled *Sundamys muelleri* was reassigned to *Rattus sp*. R3 *sensu* Pagès et al. (2010) based on *cyt b* barcoding (Table 3.1).

We also included 32 mitogenomes from *R. baluensis* (KY611359 - KY611390; Chapter 2), and other *Rattus* for which mitogenomes were available in GenBank (Australo-Papuan *Rattus*: *R. lutreolus* GU570661, *R. sordidus* GU570665, *R. praetor* NC 012461, *R. villosissimu*s NC 014864, *R. tunneyi* NC 014861, *R. leucopus* GU570659, *R. niobe* KC152486, and *R. praetor* NC_012461; Asian *Rattus*: *R. tanezumi* EU273712, *R. rattus* NC_012374, *R. nitidus* KU200226, *R. exulans* EU273711, *R. norvegicus* AJ428514, and *R. fuscipes* NC_014867). As outgroups, we included some of the closest species from the *Rattus* division: *Sundamys muelleri* KY464175, *Bandicota indica* KT029807 and *Bunomys penitus* KY464167. Additional outgroups for dating were from 6 murines belonging to 2 molecular tribes of the *Mus* branch of the phylogeny (*Apodemus chejuensis* HM034867, *A. latronum* NC_019585, *A. peninsulae* NC_016060; *Mus cervicolor* KJ530560, *M. cookii* KJ530561, *M. spretus* NC_025952; Fabre et al. 2013, 2015).

Table 3-1. Field samples and museum specimens sequenced.

Sample	*Rattus* Species	Date collected	Tissue	Elev. (m)§	Locality	Coordinates	Collector
ANSP 20348	*Rattus blangorum*	4.Apr.1939	old skin	1097	Sumatra: Mt. Leuser: Blangnanga camp	4.04,97.13	F. A. Ulmer, Jr
BOR577*	*R. exulans*	16.Mar.2013	fresh liver	357	Borneo: Sabah: Monggis substation	6.2,116.75	Miguel C.
ANSP 20309	*R. hoogerwerfi*	27.Apr.1939	old skin+dry tissue skull	2408	Sumatra: Mt. Leuser: Bivouac 5	3.87,97.13	F. A. Ulmer, Jr
ANSP 20315	*R. hoogerwerfi*	5.May.1939	old skin	2423	Sumatra: Mt. Leuser: Bivouac 6	3.87,97.15	F. A. Ulmer, Jr
ANSP 20319	*R. hoogerwerfi*	8.May.1939	old dry tissue skull	2621	Sumatra: Mt. Leuser: Bivouac 8	3.86,97.14	F. A. Ulmer, Jr
BM 19.11.5.81	*R. korinchi*	26.Apr.1914	old	2225	Sumatra: Mt. Kerinci: Sungai Kering	-1.73,101.25	Robinson and Kloss
RMNH 23151	*R. korinchi*	14.Jun.1917	old	2800	Sumatra: Mt. Talamau (=Talakmau)	0.08,99.98	E. Jacobson
NH 2147	*Rattus* sp. R3[1]	1.Feb.1980	old	16	Borneo: Sabah: Lahad Datu: Madai	4.72,118.18	-
BOR260*	*R. tanezumi*	25.Feb.2013	fresh liver	1538	Borneo: Sabah: Mt. Kinabalu: Kin. Park HQ	6.01,116.55	M.T.R. Hawkins
NH 2015	*R. tiomanicus*[2]	22.Aug.1971	old	126	Borneo: Sabah: Ulu Tuaran: Kg. Lebodon	6.15,116.37	H. Tsen
USNM 590332	*R. tiomanicus*	19.Jan.2005	fresh	22	Borneo: Sarawak: Ulu Kakas: Bukit Sarang	2.65,113.05	Helgen, K. M.
USNM 590720	*R. tiomanicus*	24.Jan.2007	fresh	22	Borneo: Sarawak: Ulu Kakas: Bukit Sarang	2.65,113.05	Helgen, K. M.

*Field code.

§Extracted from field reports, museum labels and inferred from coordinates.

[1]Originally labeled *Sundamys muelleri*, but reassigned based on *cyt b* barcoding.

[2]Originally labeled *Lenothrix canus*, but reassigned based on *cyt b* barcoding.

DNA extraction and sequencing

We extracted DNA with DNeasy Blood & Tissue Kit (Qiagen). Museum tissue samples from historical specimens were processed in an isolated ancient DNA laboratory. We constructed Illumina libraries following a double indexing protocol followed by enrichment of complete mitochondrial genomes (details in Methods on Chapters 4 and 5). Libraries were sequenced on an Illumina HiSeq 2500 with 150 PE chemistry at the Genetics Resources Core Facility at John Hopkins University.

Mitogenome assembly

We removed adaptors with cutadapt 1.8.3 (Martin 2011) using paired-end mode (-a AGATCGGAAGAGC -A AGATCGGAAGAGC -e 0.16 -m 30 -q 10 -o trimmedR1.fastq -p trimmedR2.fastq R1.fastq R2.fastq). Forward and reverse reads were paired with PEAR v0.9.6 (Zhang et al. 2014) using default parameters. The resulting assembled and unassembled forward and reverse reads were concatenated into a unique FastQ file for each library. We mapped the reads from each sample to *R. baluensis* (KY611361) with BWA 0.7.12-r1039 algorithm (Li 2013). We used SAMtools 1.3 (Li et al. 2009) to remove PCR duplicates and called consensus sequences in Geneious (http://www.geneious.com, Kearse et al. 2012) with a minimum of 2x coverage and a 75% base calling threshold. Then, we used the MAFFT v7.017 (Katoh et al. 2002) plugin in Geneious for multiple sequence alignments using the --*auto* parameter. The alignments were visually inspected and the genes were translated into amino-acids and inspected for stop codons in Geneious.

Phylogenetic reconstructions and molecular dating

We used the protein-coding genes from the mitogenomes to evaluate the evolutionary relationships between the four endemic Sundaland and other *Rattus* species in a Maximum Likelihood framework with RAxML v8.2.10 (Stamatakis 2014). Mitogenomes have been shown to provide robust support at different evolutionary depths in phylogenetic inference of Rattini (Robins et al. 2008, Wei et al. 2017; Chapters 4 and 5).

The non-protein coding genes were removed in Geneious and the gene *nd6*, which is in the light strand, was reverse-complemented. This mitogenome matrix had 57 rows, with 21 species, a length of 11339 nucleotides (69 % of the mitogenome) and 0.04 %

missing data, as calculated with AMAS (Borowiec 2016; summary -f phylip -d dna -i alignment.phy). Then, the best partition scheme was determined with PartitionFinder 2.1.1 (Lanfear et al. 2016) using the rcluster (Lanfear et al. 2014) and the RAxML algorithms. The output partition scheme was specified as input in RAxML. It arranged the data into one partition for 1st and 2nd codon position and a second with the 3rd codon position for all protein-coding genes, except codon position 2 of ATPase8. The rapid bootstrapping algorithm was run on RAxML, which converged after 350 replicates following the extended majority-rule (autoMRE) stopping criterion (raxmlHPC-PTHREADS-SSE3 -f a -m GTRGAMMA -q partitions.txt -p $RANDOM -x $RANDOM -# autoMRE -s alignment.phy -n output -T 10). The model of evolution was specified as GTR+ Γ, thus not incorporating the proportion of invariant sites (I) suggested by PartitionFinder, because the Γ parameter already considers positions evolving at low rates, thus including "I" in the model is not necessary and can further bias the estimation of both parameters reliably (Stamatakis 2016).

We also reconstructed the phylogenetic tree in a Bayesian framework with BEAST 2.4.4 (Bouckaert et al. 2014) to date the nodes. To meet the assumptions of the tree (Yule speciation process) only 1 sample per species was kept (n=21 mitogenomes). The *Rattus-Mus* split was used as a calibration point to date the tree. The final DNA matrix had 27 species, a length of 11339 nucleotides and 0.01 % of missing data. We ran PartitionFinder 2.1.1 with the greedy algorithm and branch lengths unlinked. The best scheme was used to split the alignment into 3 sets which corresponded to codon positions 1, 2 and 3, for all genes, except *nd6* codon position 3 which had its own partition. We removed it from the dataset to avoid estimating extra parameters in BEAST as a good trade-off since that region was not very informative. The alignment was then split by codon positions 1 (3,784 sites), 2 (3,779 sites) and 3 (3,611 sites), with AMAS (split -f nexus -d dna -i alignment.nex -l partitions.txt -u nexus). In BEAUTi, we set a GTR+G+I model to codon positions 1 and 3, and a HKY+G+I to codon position 2, with estimated base frequencies, as determined in PartitionFinder. We linked a relaxed clock model with frequencies sampled from a lognormal distribution to all partitions, but a relative substitution rate was estimated for each codon position. By linking clock models instead of estimating individual parameters for each gene we greatly reduced the total number of parameters to estimate by BEAST in sacrifice of integrating heterogeneity of substitution rates across mitochondrial genes in the alfa parameters of

the site models. We used the 11.81 Mya (95% CI: 11.11-12.68 Mya) suggested in Kimura et al. (2015) as a prior for the *Rattus-Mus* split. This date is based on a well-represented phylogeny of Murinae with nuclear and mitochondrial DNA with an extra calibration point from a new fossil of the *Mus-Arvicanthis* split (Fabre et al. 2013). In BEAST, the prior was specified with a lognormal distribution as suggested in Morrison (2008), and to not allow for zero values. We ran 2 chains of 50 million generations sampled every 10,000 generations. The 2 chains converged for each of the parameters in the combined log file after 10% burn-in (estimated sample sizes, ESS > 200). We generated a maximum clade credibility with TreeAnnotator after discarding the first 10% of the trees from each chain.

Mitochondrial DNA structure in the R. tiomanicus *complex*

We further evaluated the phylogenetic relationships of the closely related high elevation *R. baluensis* and the widespread, lowland *R. tiomanicus* lineages with *cyt b*, a widely-used mitochondrial marker for which there was better geographical representation of *R. tiomanicus* samples in GenBank. For *Rattus baluensis*, we extracted *cyt b* from 32 mitogenomes KY611359 - KY611390, and added *cyt b* JN675495 (Aplin et al. 2011). For *R. tiomanicus*, we included 5 samples from Thailand (KC010165- KC010168 and HM217391; Latinne et al. 2013 and Pagès et al. 2010, respectively) plus *cyt b* extracted from mitogenome KP876560 from Peninsular Malaysia, 1 sample from Java (JN675515; Aplin et al. 2011), and 6 Bornean individuals, which included USNM 590332 and 590720, from Bintulu Division, Sarawak (Table 3.1), 2 samples from Sungai Asap, Belaga, Sarawak (JF436975 and JF436986; Tamrin and Abdullah 2011), 1 from Sabah, NH 2015 (Table 3.1), and 1 from Kalimantan (JN675516; Aplin et al. 2011). We excluded other *R. tiomanicus cyt b* hits in GenBank: JF437020, because its location was missing; the Javanese sequences JN675514, EF186513 and EF186514, because they are much shorter than the rest; all sequences from Balakirev and Rozhnob (2012), because in a preliminary assessment these samples clustered with *Rattus rattus* sequences, and were collected in Vietnam, outside the native range of range of *Rattus tiomanicus*, thus they may be mis-identified specimens. A total of 33 sequences from *R. baluensis* and 16 from *R. tiomanicus* were aligned with MAFFT plugin in Geneious 8.1.5 with the --*auto* option. The alignment contained 1140 positions and 7% of missing data, as calculated with AMAS (*summary -f nexus -d dna -i cytb_alignment.nex*). A

TCS haplotype network was built in PopART (http://popart.otago.ac.nz) using 766 valid positions (with no missing data for any of the samples).

Results

Mitogenomes were successfully reconstructed from nine of the 12 individuals attempted, which covered 6 of the 7 species. Successful genomes were 96.8 to 100% complete, and had a coverage of 9.2 to 163X. Unsuccessful genomes were 0.6 to 0.7% complete and had coverage around 0.1X. The single species that was not successfully sequenced was *Rattus blangorum* (Table 3.2).

Phylogenetic relationships and molecular dating

All nodes within *Rattus* in the maximum likelihood tree and the Bayesian maximum clade credibility tree were highly supported (bootstrap support/Posterior Probability: most support values near or equal to 100/1.00, respectively; Figures 3.3 and 3.5). The four Sunda endemics (*R. hoogerwerfi, R. korinchi*, *R. tiomanicus*, and *R. baluensis*) were inside a clade within Asian-*Rattus*, well differentiated from the Australo-Papuan *Rattus*. The Sundaland endemic rats did not form a monophyletic clade. Both *Rattus tiomanicus* and *R. baluensis* are inside the *Rattus rattus* complex, which also includes the widespread Asian *R. tanezumi*, *R. rattus*, and one individual of the *Rattus* sp. lineage R3 *sensu* Pagès et al (2010). The Sumatran montane *R. korinchi* and *R. hoogerwerfi* form a well-supported clade (87/1.00) and are the closest sequenced lineages to the *Rattus rattus* group, although the branch relating it to it is very short. These Sumatran montane endemics are not sister lineages to the Bornean montane *R. baluensis*. The diversity of *R. baluensis* is nested within the diversity of *R. tiomanicus*.

Table 3-2. Information on the whole mitochondrial genome sequences generated for this study.

Code	*Rattus* Species	Coverage	Length (bp)	Mitogenome assembled (%)
ANSP 20348	*blangorum*	0.1	16311	0.6
BOR577*	*exulans*	107.4	16303	100
ANSP 20309	*hoogerwerfi*	0.13	16311	0.7
ANSP 20315	*hoogerwerfi*	0.1	16311	0.5
ANSP 20319	*hoogerwerfi*	62.7	16314	100
BM 19.11.5.81	*korinchi*	9.2	16313	96.8
RMNH 23151	*korinchi*	33.8	16312	99.8
NH 2147	sp. R3	79.2	16308	99.6
BOR260*	*tanezumi*	29	16306	100
NH 2015	*tiomanicus*	43.1	16309	100
USNM 590332	*tiomanicus*	121.4	16312	100
USNM 590720	*tiomanicus*	163	16313	100

*Field code. Specimens at the Doñana Biological Station, Spain- Not yet catalogued.

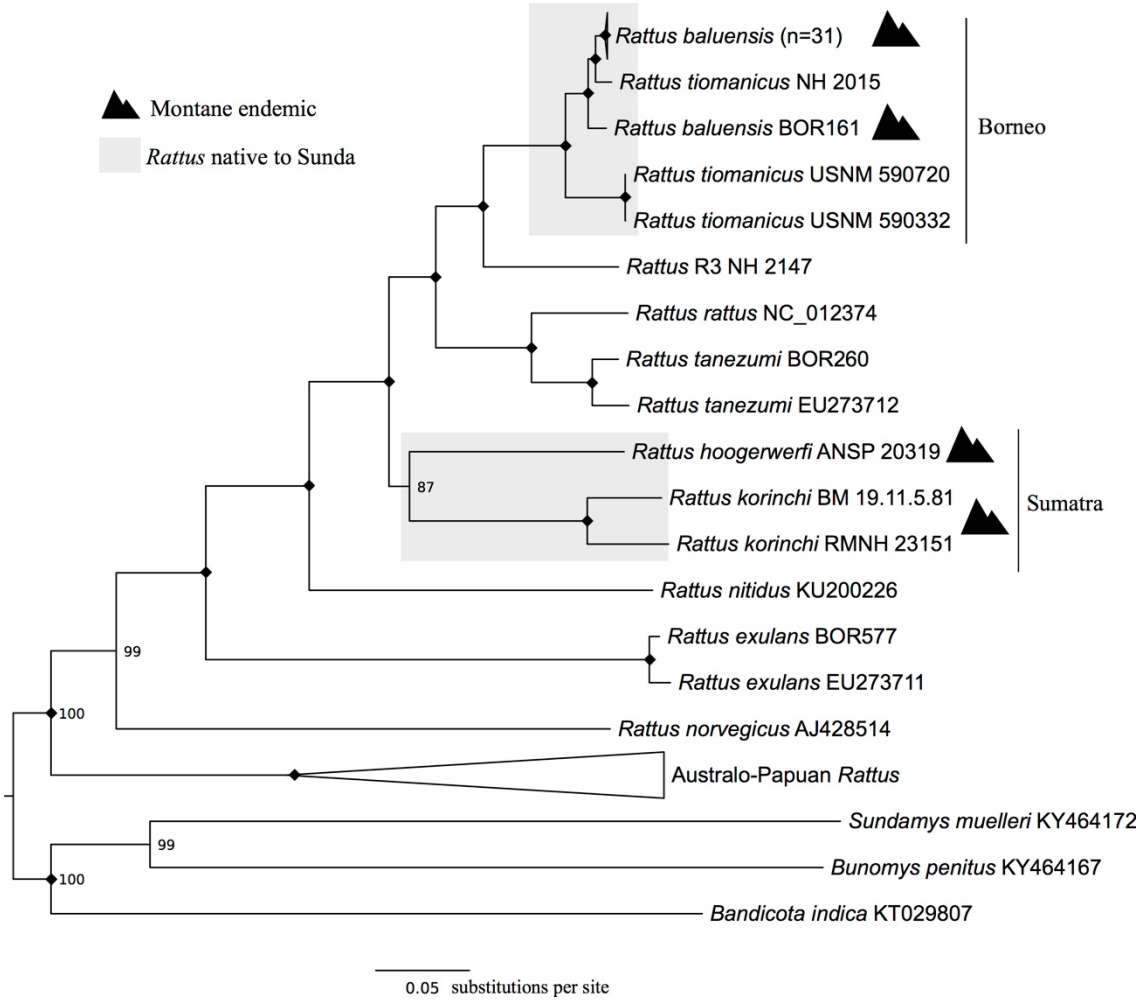


Figure 3-3. RAxML consensus tree from ML phylogenetic inference with protein-coding genes of mitogenomes. Diamonds represent 100% of bootstrap support.

According to the molecular dating, *Rattus* started to radiate about 3.3 Million years ago, Mya, (2.82-3.85), and Asian *Rattus* at 2.95 Mya (2.5-3.45). The split of the two Sumatran montane rats (*R. korinchi* and *R. hoogerwerfi*) occurred at approximately 1.3 Mya (1.04-1.51). Their divergence from its presumably closest lowland ancestor is relatively deep, 1.38 Mya (1.15-1.63), compared to the shallow 0.31 Mya (0.23-0.39) of coalescent time estimated between two of the mitogenomes from the highland *R. baluensis* and the widespread, lowland *R. tiomanicus*.

Cytochrome b diversity in Rattus tiomanicus *and* R. baluensis

A *cyt b* haplotype network showed the montane *Rattus baluensis* is nested within the greater diversity of the widespread, lowland *R. tiomanicus* (Figure 3.4; samples per haplotype in Appendix 3.1). The three *cyt b* haplotypes in *Rattus baluensis* were not monophyletic. One divergent haplotype (Rba_3) was more similar to a haplotype from *R. tiomanicus* from Sungai Asap, Belaga, Sarawak, (Rti_6, 5 mutations), than to the frequent *R. baluensis* haplotypes Rba_1 and Rba_2 (9-10 mutations; Figure 3.4).

The mitochondrial diversity in *R. baluensis* was much lower than the diversity in its sister species *R. tiomanicus*, and derives from it. Haplotypes of *R. tiomanicus* from northern Borneo were closer to *R. baluensis* than to any other Bornean or western Sunda individuals (Figure 3.3-3.4). *Rattus baluensis* was monophyletic except for one divergent haplotype, Rba_3, identified in 1 out of the 33 individuals sequenced. This haplotype had higher similarity to some haplotypes of *R. tiomanicus* than to the core diversity in *R. baluensis*. The retention of ancestral polymorphism or introgression from *R. tiomanicus* could explain this pattern. Both processes seem common at shallow evolutionary scales although and they are difficult to disentangle (Peters et al. 2007; Hailer et al. 2012; Pagès et al. 2013).

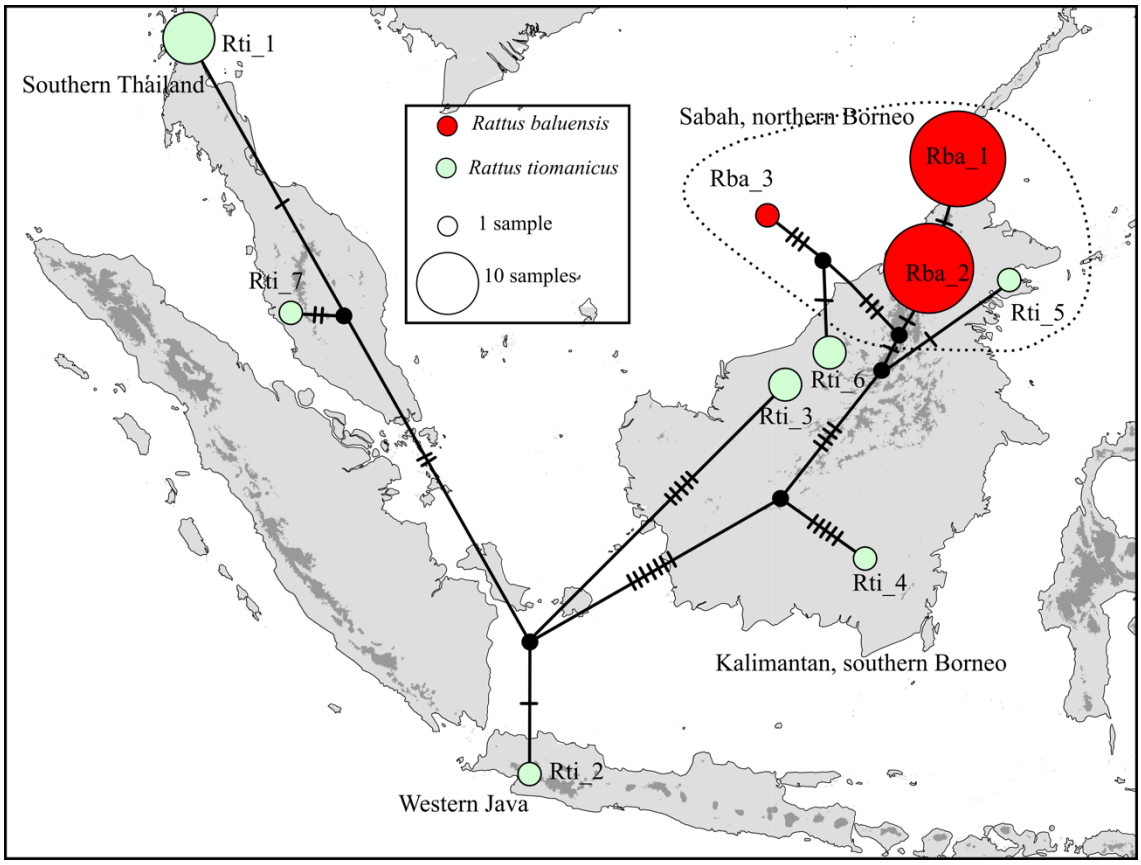


Figure 3-4. TCS haplotype network of *cyt b* sequences from *R. baluensis* and *R. tiomanicus*. The haplotypes are placed in their approximate geographic origin. A dashed line encircles all haplotypes from Sabah, northern Borneo. In the network, the circle size proportional to the number of sequences for the haplotype, black dots represent missing haplotypes, and perpendicular lines mutations between haplotypes.

Discussion

Evolutionary implications of the molecular phylogeny

According to our molecular phylogeny, the mountaintops in Borneo and Sumatra were colonized by independent *Rattus* lineages. The nested position of *R. baluensis* within the diversity of *R. tiomanicus* and the evolutionary history of other *Rattus* (Australo-Papuan *Rattus*, Rowe et al. 2011; or *Sundasciurus* squirrels, Hawkins et al. 2016) suggest polarity of lowland-to-highland divergence in Sunda mammals could common in some groups. *Rattus hoogerwerfi* and *R. korinchi* are sister taxa which coalesce in relative long branches of the tree, with a common ancestor at around 1.3 Mya, and they are peripheral to the *Rattus rattus* complex. These results resolve the uncertain taxonomic position (see Musser and Carleton 2005) of these high-elevation Sumatran and are consistent with the taxonomic placement after morphological descriptions in

Musser and Newcomb (1983). Their deep divergence within Asian *Rattus* contrasts with the young origin of *Rattus baluensis* from *Rattus tiomanicus* (<0.39 Mya).

Origin of Rattus baluensis

We found the low mitochondrial diversity in *R. baluensis* derives from the local diversity of *R. tiomanicus*. These two species have not reached reciprocal monophyly. The genetic structure of *R. baluensis* with respect to *R. tiomanicus* illustrates the predicted genetic consequences vicariance processes produce among populations with different sizes, in which the smaller population (*R. baluensis* in this case) will become monophyletic first, while the larger one (*R. tiomanicus*) will remain paraphyletic for some longer time before reaching reciprocal monophyly (Zink and Barrowclough 2008). The widespread *R. tiomanicus* is restricted to lowlands or mid- elevations across Sundaland, and no evidence hints at the upper slopes of Mt. Kinabalu being part of its ancestral distribution. A more likely scenario invokes a founder effect in which a local lineage of *R. tiomanicus* colonized mountain habitat on Mt. Kinabalu. This makes *R. baluensis* another example of a Mt. Kinabalu endemic which originated from a lowland taxon (Merckx et al. 2015), and it is also consistent with the low genetic diversity reported in *R. baluensis*.

Founder effects have been widely described in mammals arriving to islands. One common founder effect is the loss in genetic diversity in the island populations as compared to mainland relatives, since only a fraction of a species' diversity would arrive (Berry 1996; Frankham 1997; Abdelkrim et el. 2005). An analogous process could happen on 'islands' of montane forests surrounded by tropical lowlands, although we have found no examples in the literature. Often, founder effects are linked to rapid morphological divergence, as has been described for invasive *Rattus* (Patton et al. 1975; Pergams et al. 2015), or other rodents (Berry 1996; Pergams and Ashley 2009), but linking the consequence to the cause, "founder effect", is challenging (Pergams and Lacy 2008; Yeung et al. 2011). The mechanisms behind this rapid morphological divergence are debated. Mayr (1954) introduced the term "genetic revolution" in which founder events cause allele frequencies to change rapidly because their adaptive value is dependent on the genetic background (=allelic combinations) of the few founders, plus there will be strong selection on newly exposed alleles in homozygosis. Different theories explain how founder events drive a population to a new local fitness peak with no need of environmental change (reviews in Barton and Charlesworth 1984;

Templeton 2008). In the case of *R. baluensis* this process could have been boosted by change in selection pressures imposed by the new, montane habitat (colder, less oxygen) or biotic interactions (e.g. relaxation of predation or parasitism).

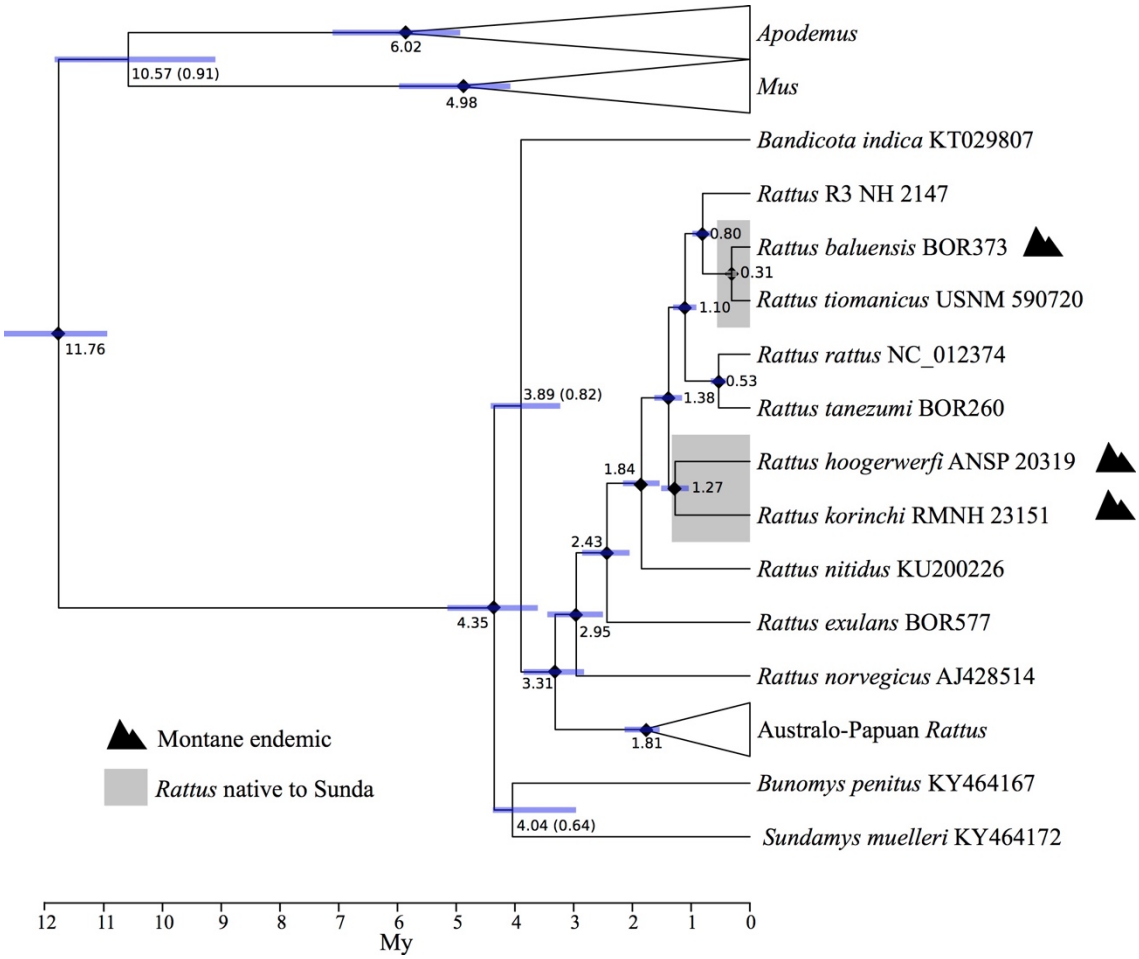


Figure 3-5. Maximum clade credibility tree from BEAST analysis using protein-coding genes of mitogenomes. Node ages in millions of years ago (Mya) with their 95 % HPD are represented each node. Diamonds represent *PP* = 1.00. *PP* below 1.00 are indicated in parenthesis.

Evidence of morphological convergence in montane Rattus

Few characters differentiate the skulls of *R. baluensis* and *R. tiomanicus* (frequency of occurrence of cusps t3 in first and second molars and size of skull, larger in *R. baluensis*; Musser 1986). This is a low level of interspecific differentiation among congeneric Rattini (Musser and Newcomb 1983; Musser 1986). Little divergence in skulls contrasts with marked differences in external morphology between these two species. The differences in appearance are similar to those observed between the montane *R. korinchi* and *R. hoogerwerfi* and other lowland *Rattus*. The thick dark fur and larger skull and body measurements have been key traits for the morphological

definition Sunda montane *Rattus*, but they likely reflect convergent adaptation to mountain habitats rather than common ancestry, as our molecular phylogeny revealed. This kind of morphological convergence is not unique to *Rattus*. For instance, it also hindered the taxonomic position of the co-distributed ground squirrel *Dremomys everetti*, which was recently moved to *Sundasciurus everetti* (Hawkins et al. 2016). These traits seems to be also shared by other co-distributed montane small mammals such as *Maxomys hylomyoides* or *Sundamys infraluteus*.

The larger size of *R. baluensis* as opposed to its lowland sister species *R. tiomanicus* (mean ± sd in mm for *R. baluensis*/*R. tiomanicus*; head body, HB: 170±8.4 for n=23/157.8±14.6 for n=5; greatest length of skull, GLS: 40.8±1.3 for n=24/37.4±1.3 for n=12; from Musser 1986, and Musser and Califia 1982) may suggest an "island" effect on its reduced mountain habitat. Particularly in rodents, there is a general negative correlation between island size and body size, probably as a convergence to a better physiological efficiency, which is allowed by reduced predatory and competitive interspecific interactions, while food availability does not become a limiting factor for small mammals at these small areas (Heaney 1978; Lomolino 1985). This same pattern is particularly marked in the larger insular populations of *R. tiomanicus* in many islands of eastern Borneo (most GLS around 40 mm vs 37.4 mm in mainland Borneo; Musser 1986; Musser and Califia 1982). This pattern could also be extended to the even larger-bodied Sumatran *R. hoogerwerfi* (HB: 182.7±6.7 for n=20; GLS: 42.9±0.8 for n=16) and *R. korinchi* (HB: 166 and 169; GLS: 41 and 41.8) which inhabit very restricted high altitude islands. Alternatively or additionally, this could be an example of Bergmann's rule, which describes a pattern of larger body size associated with colder climate.

The darker pelage in the montane *R. baluensis*, *R. hoogerwerfi*, and *R. korinchi*, compared to other lowland *Rattus* could respond to the Gloger's rule (Gloger 1833), initially proposed for explaining melanism of birds in the wet tropics. In mammals, it has also been described, but its adaptive value is debated, and could be related to thermoregulation, camouflage or resistance to bacterial degradation of the hair in moist habitats (Lai et al. 2008; Kamilar and Bradley 2011). The evaluation of this effect could be difficult to disentangle from a potential island effect. Greater melanism is not unique to mountain lineages, as it is also common in Sunda mammals on islands, such as *R. tiomanicus* from the Maratua Islands (Musser and Califia 1982) or some species from the Mentawai Islands (Corbet and Hill 1992). Completely different mechanisms could

explain the darker fur in these highland lineages due to the pleiotropic effect of melanocortins. In vertebrates, melanism has been shown to covariate with phenotypical traits, such as aggressiveness (Ducrest et al. 2008), which could be positively selected for on islands where interspecific competition is greater, such as has been suggested in birds (Albert and Vargas-Castro 2015). The higher density of individual *R. baluensis* and other co-distributed highland species as compared to their adjacent lowland relatives may support this (Hawkins 2015).

Taken together the ecological, morphological and genetic data suggests *R. baluensis* and *R. tiomanicus* are 2 independent evolutionary units that diverged very recently and still could fall the "gray zone" along the speciation process in which alternative species concepts can come into conflict (de Queiroz 2007). In this context, *Rattus baluensis* offers a particularly good system to study speciation genomics in the light of founder events and/or new selection pressures, and to adaptation to high altitudes (Cheviron and Brumfield 2012).

Acknowledgements

We thank the curators and managers that facilitated access to different mammal collections: N. Gilmore and T. Daeshler, ANSP, Philadelphia; K. Helgen, N. Edminson, and the Mammal Division at the NMNH, Washington DC; S. van der Mije and P. Kamminga, NBC, Leiden; R. Portela, NHM, London; A. Lo, Sabah Museum, Kota Kinabalu. S.Y.W. Ho provided valuable insight regarding molecular dating. Logistical support was provided by Laboratorio de Ecología Molecular, Estación Biológica de Doñana, CSIC (LEM-EBD). We also thank Sabah Parks for research permits (TS/PTD/5/4 Jld. 45 (33) and TS/PTD/5/4 Jld. 47 (25)) and various kind of support, the Economic Planning Unit (reference: 100-24/1/299), and export permits from the Sabah Wildlife Department (JHL.600-3/7 Jld.7/19 and JHL.600-3/7 Jld.8/) and Sabah Biodiversity Council (Ref: TK/PP:8/8Jld.2). MCS received support from the SYNTHESYS Project (http://www.synthesys.info/) which is financed by the European Community Research Infrastructure Action under the FP7 Integrating Activities Program: SYNTHESYS ACCESS NL-TAF-5588 to NBC, Leiden, and GB-TAF-5303 to NHM, London. The Spanish Ministry of Science and Innovation grants CGL2010-21524 and CGL2014-58793-P also supported this work. MCS is supported by the Spanish Ministry of Science and Innovation Predoctoral Fellowship BES-2011-049186.

Literature cited

ABDELKRIM, J., M. PASCAL AND S. SAMADI. 2005. Island colonization and founder effects: The invasion of the Guadeloupe islands by ship rats (*Rattus rattus*). Molecular Ecology 14:2923–2931.

ALBERT, J., C. UY AND L. E. VARGAS-CASTRO. 2015. Island size predicts the frequency of melanic birds in the color-polymorphic flycatcher *Monarcha castaneiventris* of the Solomon Islands. The Auk 132:787–794.

APLIN, K. P., H. SUZUKI, A.A. CHINEN, R.T. CHESSER, J. TEN HAVE, S.C. DONNELLAN, J. AUSTIN, A. FROST, J.P. GONZALEZ, V. HERBRETEAU, AND F. CATZEFLIS. 2011. Multiple Geographic Origins of Commensalism and Complex Dispersal History of Black Rats. PLoS ONE 6:e26357.

BALAKIREV, A. A. E. AND V. V. ROZHNOV. 2012. Contribution to the species composition and taxonomic status of some *Rattus* inhabiting Southern Vietnam and Sundaland. Russian Journal of Theriology 11:33–45.

BARTON, N. H. AND B. CHARLESWORTH. 1984. Genetic Revolutions, Founder Effects, and Speciation. Annual Review of Ecology and Systematics 15:133–164.

BEALL, C.M., G.L. CAVALLERI, L. DENG, R.C. ELSTON, Y. GAO, J. KNIGHT, C. LI, J.C. LI, Y. LIANG, M. MCCORMACK AND H.E. MONTGOMERY. 2010. Natural selection on EPAS1 (HIF2α) associated with low hemoglobin concentration in Tibetan highlanders. Proceedings of the National Academy of Sciences, 107: 11459-11464.

BERRY, R. J. 1996. Small mammal differentiation on islands. Philosophical Transactions of the Royal Society of London. Series B. 351:753–64.

BOROWIEC, M. L. 2016. AMAS: a fast tool for alignment manipulation and computing of summary statistics. PeerJ 4:e1660.

BOUCKAERT, R., J. HELED, D. KÜHNERT, T. VAUGHAN, C.H. WU, D. XIE, M.A. SUCHARD, A. RAMBAUT AND A.J. DRUMMOND. 2014. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. PLoS Computational Biology 10: e1003537.

CHASEN, F. N. 1939. Two new mammals from North Sumatra. Treubia 17:207–208.

CHEVIRON, Z. A., AND R. T. BRUMFIELD. 2012. Genomic insights into adaptation to high-altitude environments. Heredіy 108: 354-361.

CHEVIRON, Z.A., M.D. CARLING, AND R.T. BRUMFIELD. 2011. Effects of postmortem interval and preservation method on RNA isolated from field-preserved avian tissues. The Condor 113: 483-489.

CHEVIRON, Z. A., A. D. CONNATY, G. B. MCCLELLAND AND J. F. STORZ. 2013. Functional genomics of adaptation to hypoxic cold-stress in high-altitude deer mice: transcriptomic plasticity and thermogenic performance. Evolution 68:48–62.

CORBET, G. B. AND J. E. HILL. 1992. The mammals of the Indomalayan region: a systematic review. P. in. Oxford University Press.

DE QUEIROZ, K. DE. 2007. Species concepts and species delimitation. Systematic Botany 56:879–886.

DUCREST, A. L., L. KELLER AND A. ROULIN. 2008. Pleiotropy in the melanocortin system, coloration and behavioural syndromes. Trends in Ecology and Evolution 23:502–510.

FABRE, P. FABRE, P.H., MUSSER, G.G., FITRIANA, Y.S., FJELDSÅ, J., JENNINGS, A., JØNSSON, K.A., J. KENNEDY, J. MICHAUX, G. SEMIADI, N. SUPRIATNA AND K.M. HELGEN. 2013. A new genus of rodent from Wallacea (Rodentia: Muridae: Murinae: Rattini), and its implication for

biogeography and Indo-Pacific Rattini systematics. Zoological Journal of the Linnean Society 169:408–447.

FABRE, P.H., Y. CHAVAL, A. MORTELLITI, V. NICOLAS, K. WELLS, J.R. MICHAUX, AND V. LAZZARI. 2015. Molecular phylogeny of South-East Asian arboreal murine rodents. Zoologica Scripta 45:349–364.

FRANKHAM, R. 1997. Do island populations have less genetic variation than mainland populations? Heredity 78: 311–327.

GLOGER, C. 1833. Das Abandern der Vogel durch Einfluss des Klimas. August Schulz, Breslau, Germany.

GORKHALI, N.A., K. DONG, M. YANG, S. SONG, A. KADER, B.S. SHRESTHA, X. HE, Q. ZHAO, Y. PU, X. LI AND J. KIJAS. 2016. Genomic analysis identified a potential novel molecular mechanism for high-altitude adaptation in sheep at the Himalayas. Scientific Reports 6: 29963.

HAILER, F. V.E. KUTSCHERA, B.M. HALLSTRÖM, D. KLASSERT, S.R. FAIN, J.A. LEONARD, U. ARNASON AND A. JANKE. 2012. Nuclear genomic sequences reveal that polar bears are an old and distinct bear lineage. Science 336: 344–347.

HAWKINS, M. T. R. 2015. Biogeography and Phylogeography of Mammals of Southeast Asia: A Comparative Analysis Utilizing Macro and Microevolution. Doctoral thesis. George Mason University.

HAWKINS, M. T. R., K. M. HELGEN, J. E. MALDONADO, L. L. ROCKWOOD, M. T. N. TSUCHIYA AND J. A. LEONARD. 2016. Phylogeny, biogeography and systematic revision of plain long-nosed squirrels, (genus *Dremomys*, Nannosciurinae). Molecular Phylogenetics and Evolution 94:752–764.

HEANEY, L. R. 1978. Island Area and Body Size of Insular Mammals: Evidence from the Tri-Colored Squirrel (*Callosciurus prevostii*) of Southeast Asia. Evolution 32:29–44.

IUCN. 2015. The IUCN Red List of threatened species. Ver. 2015.3 (www.iucnredlist.org). Accessed 16 December 2015.

KAMILAR, J. M. AND B. J. BRADLEY. 2011. Interspecific variation in primate coat colour supports Gloger's rule. Journal of Biogeography 38:2270–2277.

KATOH, K., K. MISAWA, K. KUMA AND T. MIYATA. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research 30:3059–3066.

KEARSE, M., R. MOIR, A. WILSON, S. STONES-HAVAS, M. CHEUNG, S. STURROCK, S. BUXTON, A. COOPER, S. MARKOWITZ, C. DURAN AND T. THIERER. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28:1647–1649.

KIMURA, Y., M.T. HAWKINS, M.M. MCDONOUGH, L.L. JACOBS AND L.J. FLYNN. 2015. Corrected placement of *Mus-Rattus* fossil calibration forces precision in the molecular tree of rodents. Scientific Reports 5:14444.

LAI, Y. C., T. SHIROISHI, K. MORIWAKI, M. MOTOKAWA AND H. T. YU. 2008. Variation of coat color in house mice throughout Asia. Journal of Zoology 274:270–276.

LANFEAR, R., B. CALCOTT, D. KAINER, C. MAYER AND A. STAMATAKIS. 2014. Selecting optimal partitioning schemes for phylogenomic datasets. BMC Evolutionary Biology 14:82.

LANFEAR, R., P. B. FRANDSEN, A. M. WRIGHT, T. SENFELD AND B. CALCOTT. 2016. PartitionFinder 2: New Methods for Selecting Partitioned Models of Evolution for Molecular and Morphological Phylogenetic Analyses. Molecular Biology and Evolution 34:772-773.

LATINNE, A., S. WAENGSOTHORN, P. ROJANADILOK, K. EIAMAMPAI, K. SRIBUAROD AND J. R. MICHAUX. 2013. Diversity and endemism of Murinae rodents in Thai limestone karsts. Systematics and Biodiversity 11:323–344.

LI, H., B. HANDSAKER, A. WYSOKER, T. FENNELL, J. RUAN, N. HOMER, G. MARTH, G. ABECASIS AND R. DURBIN. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.

Li, H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:1303.3997.

LOMOLINO, M. V. 1985. Body Size of Mammals on Islands: The Island Rule Reexamined. The American Naturalist 125:310.

MARICIC, T., M. WHITTEN AND S. PÄÄBO. 2010. Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. PLoS One 5:e14004.

MARTIN, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17:10–12.

MAYR, E. 1954. Change of genetic environment and evolution. In: Evolution as a Process, ed. HuxleyJ, Hardy A, Ford E. London: Allen and Unwin.

MERCKX, V. S. F. T. ET AL. 2015. Evolution of endemism on a young tropical mountain. Nature 524:347–350.

MEYER, M., AND M. KIRCHER. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harbor Protocols 2010:pdb.prot5448.

MILLER, G. S. 1942. Zoological Results of the George Vanderbilt Sumatran Expedition, 1936-1939. Part V: Mammals Collected by Frederick A. Ulmer, Jr. on Sumatra and Nias. Proceedings of the National Academy of Sciences of Philadelphia 94:107–165.

MORRISON, D. A. 2008. How to summarize estimates of ancestral divergence times. Evolutionary Bioinformatics 2008:75–95.

MUSSER, G. G. 1986. Sundaic *Rattus*: definitions of *Rattus baluensis* and *Rattus korinchi*. American Museum Novitates 2862:1–24.

MUSSER, G. G. AND D. CALIFIA. 1982. Identities of rats from Pulau Maratua and other islands off East Borneo. American Museum Novitates:1–30.

MUSSER, G. G. AND M. D. CARLETON. 2005. Superfamily Muroidea. Pp. 894–1531inMammal Species of the World: a taxonomic and geographic reference (D. E. Wilson & D. M. Reeder, eds.). 3rd edition. The Johns Hopkins University Press, Baltimore.

MUSSER, G. G. AND C. NEWCOMB. 1983. Malaysian murids and the giant rat from Sumatra. Bulletin of the American Museum of Natural History 174: 327–598.

NOR, S. 2001. Elevational diversity patterns of small mammals on Mount Kinabalu, Sabah, Malaysia. Global Ecology and Biogeography 10: 41-62.

PAGÈS, M., Y. CHAVAL, V. HERBRETEAU, S. WAENGSOTHORN, J.F. COSSON, J.P. HUGOT, S. MORAND AND J. MICHAUX. 2010. Revisiting the taxonomy of the Rattini tribe: a phylogeny-based delimitation of species boundaries. BMC Evolutionary Biology 10: 184.

PAGÈS, M., M. GALAN, Y. CHAVAL, J. CLAUDE, J. MICHAUX, S. PIRY, S. MORAND AND J.F. COSSON. 2013. Cytonuclear discordance among Southeast Asian black rats (*Rattus rattus* complex). Molecular Ecology 22:1019–34.

PATTON, J. L., S. Y. YANG AND P. MYERS. 1975. Genetic and morphological divergence among introduced rat populations (*Rattus rattus*) of the Galapagos archipelago. Systematic Zoology 24:296–310.

PERGAMS, O. R. W. AND M. V ASHLEY. 2009. Rapid Morphological Change in Channel Island Deer Mice. Evolution 53:1573–1581.

PERGAMS, O. R. W., D. BYRN, K. L. Y. LEE AND R. JACKSON. 2015. Rapid morphological change in black rats (*Rattus rattus*) after an island introduction. PeerJ 3:e812.

PERGAMS, O. R. W. AND R. C. LACY. 2008. Rapid morphological and genetic change in Chicago-area *Peromyscus*. Molecular Ecology 17:450–463.

PETERS, J. L., Y. ZHURAVLEV, I. FEFELOV, A. LOGIE AND K. E. OMLAND. 2007. Nuclear loci and coalescent methods support ancient hybridization as cause of mitochondrial paraphyly between gadwall and falcated duck (*Anas* spp.). Evolution 61:1992–2006.

ROBINS, J. H., P. A. MCLENACHAN, M. J. PHILLIPS, L. CRAIG, H. A. ROSS AND E. MATISOO-SMITH. 2008. Dating of divergences within the *Rattus* genus phylogeny using whole mitochondrial genomes. Molecular Phylogenetics and Evolution 49:460–6.

ROBINSON, H. AND C. KLOSS. 1918. Results of an expedition to Korinchi Peak, 12,400 ft., Sumatra. 1. Mammals. Journal of the Federated Malay State Museums 8:1–80.

ROBINSON, H. AND C. KLOSS. 1919. On Mammals chiefly from the Ophir district, West Sumatra. Journal of the Federated Malay State Museums 7:299–323.

ROWE, K.C., K.P. APLIN, P.R. BAVERSTOCK AND C. MORITZ. 2011. Recent and rapid speciation with limited morphological disparity in the genus *Rattus*. Systematic Biology, 60: 188-203.

SODY, H. J. V. 1941. On a collection of rats from the Indo-Malayan and Indo-Australian regions. Treubia 18:255–325.

STAMATAKIS, A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.

STAMATAKIS, A. 2016. The RAxML v8.0.X Manual.

STORZ, J. F. 2007. Hemoglobin function and physiological adaptation to hypoxia in high-altitude mammals. Journal of Mammalogy, 88: 24-31.

STORZ, J. F., A.M. RUNCK, S.J. SABATINO, J.K. KELLY, N. FERRAND, H. MORIYAMA, R.E. WEBER AND A. FAGO. 2009. Evolutionary and functional insights into the mechanism underlying high-altitude adaptation of deer mouse hemoglobin. Proceedings of the National Academy of Sciences of the United States of America 106: 14450–5.

TAMRIN, N. A. AND M. T. ABDULLAH. 2011. Molecular phylogenetics and systematics of five genera of Malaysian inferred from partial mitochondrial cytochrome c oxidase subunit I (COI) gene. Journal of Science and Technology in the Tropics 7:75–86.

TEMPLETON, A. R. 2008. The reality and importance of founder speciation in evolution. BioEssays 30:470–9.

WASSERMAN, D. AND D.J. NASH. 1979. Variation in body size, hair length, and hair density in the deer mouse *Peromyscus maniculatus* along an altitudinal gradient. Ecography 2:115-118.

WEI, H., F. LI, X. WANG, Q. WANG, G. CHEN, H. ZONG AND S. CHEN. 2017. The characterization of complete mitochondrial genome and phylogenetic relationship within *Rattus* genus (Rodentia: Muridae). Biochemical Systematics and Ecology 71: 179–186.

YEUNG, C. K., P.W. TSAI, R.T. CHESSER, R.C. LIN, C.T. YAO, X.H. TIAN AND S.H. LI. 2011. Testing founder effect speciation: Divergence population genetics of the Spoonbills *Platalea regia* and *Pl. minor* (Threskiornithidae, Aves). Molecular Biology and Evolution 28:473–482.

ZHANG, J., K. KOBERT, T. FLOURI AND A. STAMATAKIS. 2014. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics 30:614–620.

ZINK, R. M. AND G. F. BARROWCLOUGH. 2008. Mitochondrial DNA under siege in avian phylogeography. Molecular Ecology 17:2107–2121.

Appendix 3.1 Haplotype assignation

Specimens assigned to each *cytochrome b* haplotype for *Rattus baluensis* and *R. tiomanicus* in the haplotype network reconstruction in Figure 3.4. Codes in Tables 2.1 and 3.1.

Rba_1: B0993 BOR201, BOR209, BOR210, BOR212, BOR216, BOR223, BOR230, BOR326, BOR344, BOR529, BOR533, BOR362, BOR373, BOR384, BOR391, S0903.

Rba_2: BOR207, BOR354, BOR519, BOR528, BOR532, BOR540, BOR548, BOR343, BOR348, BOR383, BOR392, BOR393, BOR398, BOR399, JN675495.

Rba_3: BOR161.

Rti_1: HM217391, KC010165, KC010166, KC010167, KC010168.

Rti_2: JN675515.

Rti_3: USNM 590332, USNM 590720.

Rti_4: JN675516.

Rti_5: NH 2015.

Rti_6: JF436975, JF436986.

Rti_7: KP876560.

Chapter 4 The generic status of *Rattus annandalei* Bonhote, 1903 (Rodentia, Murinae) and its evolutionary implications

Miguel Camacho Sanchez[1], Jennifer A. Leonard[1], Yuli Fitriana[2], Marie-Ka Tilak[3], and Pierre-Henri Fabre[3,4]

[1]Conservation and Evolutionary Genetics Group, Estación Biológica de Doñana (EBD-CSIC), Avda Américo Vespucio sn, 41092, Sevilla, Spain.

[2]Museum Zoologicum Bogoriense, Research Center for Biology, Indonesian Institute of Sciences (LIPI), Jl. Raya Jakarta-Bogor Km.46 Cibinong 16911, Indonesia.

[3]Institut des Sciences de l'Évolution (ISEM, UMR 5554 CNRS), Université Montpellier II, Place E. Bataillon - CC 064 - 34095 Montpellier Cedex 5, France.

[4]National Museum of Natural History, Smithsonian. Institution, P.O. Box 37012, MRC 108, Washington, DC 20013-7012, USA.

Chapter 4: Taxonomy of *Rattus annandalei*

Abstract

The taxonomic position of Annandale's rat (*Rattus annandalei* Bonhote, 1903) has been uncertain given its mix of *Rattus*-like and *Sundamys*-like morphological features. Annandale's rat and all described species in *Sundamys* (the lowland *S. muelleri,* and the montane *S. maxi* and *S. infraluteus*) are endemic to Sundaland, a center of diversification and endemism for their tribe, the Rattini. Using mitochondrial genomes and 3 nuclear markers (*rag1*, *rbp3*, *ghr*) we provide the first phylogenetic framework for *Sundamys*. We find that *Rattus annandalei* is nested within *Sundamys*, and that the 4 species likely diverged during the Pleistocene. We move *Rattus annandalei* to *Sundamys* and provide an emended diagnosis for *Sundamys*. Using geometric morphometric analyses of skulls and mandibles, we identify morphological differences between lowland and highland species of *Sundamys* that may be associated with adaptations to distinct diets.

Introduction

Annandale's rat, *Rattus annandalei* Bonhote 1903, is 1 of the 5 *Rattus* species endemic to Sundaland (also: *R. tiomanicus* complex, *R. baluensis*, *R. hoogerwerfi*, and *R. korinchi*). It is a lowland species restricted to southern Peninsular Malaysia, Singapore, eastern Sumatra, and the islands of Padang and Rupat (Figure 4.1). Its proper taxonomic affiliation is uncertain due to the presence of morphological characters associated with both *Rattus* and *Sundamys* (Musser and Newcomb 1983; Musser and Carleton 2005). In the 1st description of *Rattus annandalei*, Bonhote (1903) highlighted its large bullae, which contrasted with the small bullae in *Sundamys* (Musser and Newcomb 1983). However, a shared 2n = 42 chromosomes (Yong 1969; Yosida 1973) and similar allozyme profiles (Chan 1977; Chan et al. 1979) pointed to a potential affinity with *Sundamys*.

Sundamys is endemic to Sundaland (Figure 4.1, Musser and Newcomb 1983; IUCN 2015). It includes a widespread, lowland species, *S. muelleri*, and 2 lineages restricted to mountain ranges, *S. maxi* (Java) and *S. infraluteus* (Sumatra and Borneo). The lowland *S. muelleri* inhabits all major islands of Sundaland except Java, and is sympatric with *R. annandalei* in eastern Sumatra and the southern Malay Peninsula (Musser and Newcomb 1983). *Sundamys muelleri* and *R. annandalei* occur in a variety of similar lowland habitats (Harrison and Lim 1950; Harrison 1955; Lim 1966, 1970; Muul and Liat 1971; Wilson et al. 2006). Despite being an abundant lowland species, there are occasional records of *S. muelleri* in montane forest up to 1800 m (Musser and Newcomb 1983). The mountain species *S. maxi* is only known from 21 specimens collected from 1932 to 1935 between 900 and 1350 m from 2 mountain locations around 58 km apart, Tjiboeni (*cf.* Cibuni) and Mount Gede Pangrango, in western Java (Musser and Newcomb 1983). *Sundamys infraluteus* occurs in montane habitats in the north of Borneo and along several mountain ranges in Sumatra, at different elevations ranging from 700 to 2400 m, in habitats such as lower montane oak forest and mossy forest (Musser and Newcomb 1983; Musser and Carleton 2005; Cranbrook et al. 2014). Mountain habitats in Sundaland may be associated with morphological convergence in some mammals, such as has been shown with the skull of the Bornean mountain ground squirrel, *Sundasciurus everetti* (Hawkins et al. 2016) or the external morphology of Sunda highland *Rattus* (Musser 1986). The elevational distribution of the *Sundamys*

species allow us to explore possible morphological divergence associated with lowland or mountain habitats.

We use protein-coding mitochondrial genes and 3 nuclear loci (*rag1*, *rpb3*, and *ghr*) to determine the relationship between *R. annandalei* and all recognized species of *Sundamys,* along with representative species of *Rattus*. We find that *R. annandalei* is phylogenetically placed within *Sundamys,* and we identify morphological characters that define this group.

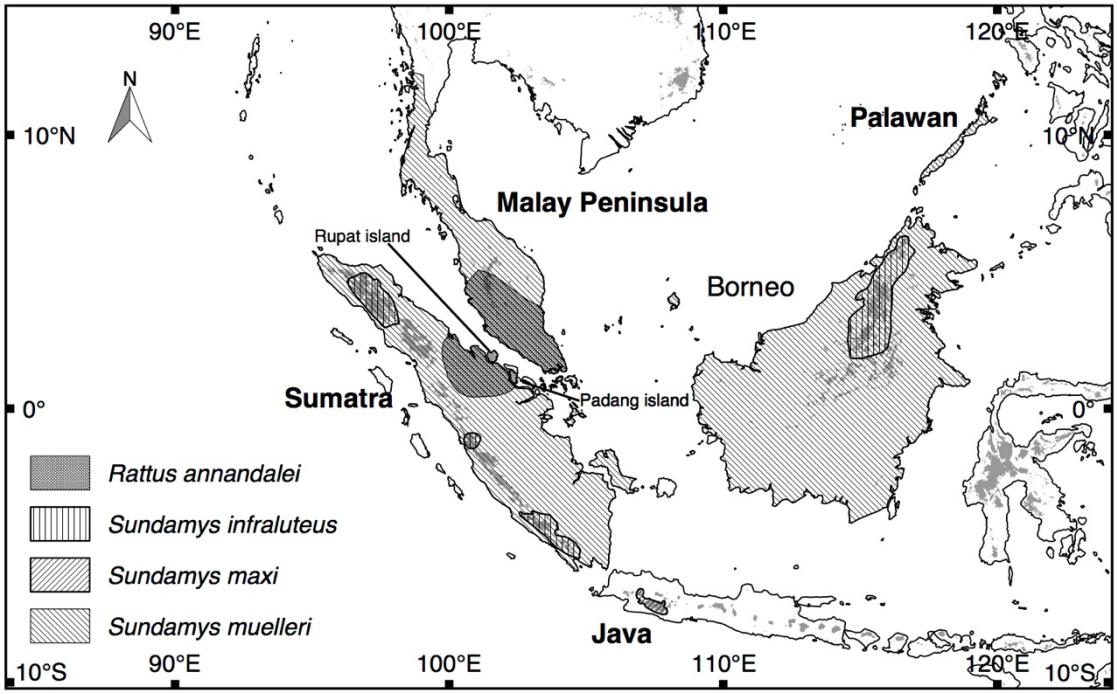


Figure 4-1. Distribution of *Rattus annandalei* and the 3 recognized species of *Sundamys*. Shaded areas indicate zones above 1,000 m (data from IUCN 2015).

Materials and Methods

Molecular taxon and gene sampling

We sampled a total of 27 species. Ingroup taxa included *R. annandalei* and species from the 2 genera with which it has morphological affinities: *Sundamys* and *Rattus*. In total, we included *R. annandalei*, all species in *Sundamys* (*S. infraluteus*, *S. maxi*, and *S. muelleri*), 8 Australo-Papuan and 4 Asian *Rattus* (*R. praetor*, *R. niobe*, *R. leucopus*, *R. tunneyi*, *R. villosissimus*, *R. sordidus*, *R. lutreolus*, and *R. fuscipes*; and *R. norvegicus*, *R. tanezumi*, *R. rattus*, and *R. exulans*, respectively), and other Rattini with close affinities (*Bandicota indica*, *Berylmys berdmorei, Halmaheramys bokimekot*, and

Paruromys dominator). We also included several outgroups from the Maxomys division (*Maxomys surifer*), Dacnomys division (*Niviventer confucianus*, *Niviventer excelsior*, *Lenothrix canus*, and *Leopoldamys edwardsi*), and Micromys division (*Micromys minutus*). We analyzed 1 sequence per species except for the 3 *Sundamys* species and *R. annandalei*, for which we sequenced 2 individuals per species (Table 4.1).

When available, data were downloaded from GenBank (Table 4.1). Additionally, data were collected from tissue samples of historical museum specimens, modern tissue samples from animals collected in the field and vouchered (all museum acronyms in *Geometric morphometric procedures*), or from non-vouchered (NV) individuals that were sampled and released in the field. These included modern tissue of vouchered *Rattus annandalei* MZB 28969 and 28971 from Sumatra, tissue of historical specimens *Sundamys maxi* RMNH 21479 and 14208 from Java, modern samples of *Sundamys muelleri* EBD 30384M and BOR448 (NV), *S. infraluteus* (field codes BOR251 and BOR510, EBD, not yet cataloged), and *Lenothrix canus* (field code BOR036, EBD, not yet cataloged) (Kinabalu National Park, Malaysia); *Leopoldamys edwardsi* (CBGP R4222)*, Berylmys berdmorei* (CBGP L0006) and *Maxomys surifer* (CBGP R4223) (Thailand); *Halmaheramys bokimekot* (MZB 33262) (Halmahera); *Bunomys penitus* (field code MORT_SP, NV) and *Paruromys dominator* (field code MORT_S46, NV) (Sulawesi). Samples we collected were taken according to the guidelines of the American Society of Mammalogists (Sikes et al. 2011), and as approved by institutional animal care and use committees (Estación Biológica de Doñana Proposal Number CGL2010-21524).

We targeted mitogenomes and 3 nuclear loci previously found to be informative in murine phylogenies: *rag1* (recombination activating gene 1, exon 1); *rbp3* (retinol-binding protein 3, exon 1); and *ghr* (growth hormone receptor, exon 10) (Steppan et al. 2004, 2005; Jansa et al. 2006; Lecompte et al. 2008; Rowe et al. 2008, 2011; Pagès et al. 2010; Fabre et al. 2013; Schenk et al. 2013).

Table 4-1. GenBank accession numbers for sequences used for phylogenetic reconstructions. Sequences generated in this study indicated in bold.

Species	Mitochondria	*rbp3*	*ghr*	*rag1*
Bandicota indica	KT029807[1]	HM217646[2]	-	-
Bunomys penitus	**KY464167**	KC878202[3]	KC878171[3]	-
Halmaheramys bokimekot	**KY464168**	KF164256[4]	KF164271[4]	-
Paruromys dominator	**KY464169**	KC953433[5]	EU349822[6]	KJ607320
Rattus exulans	KJ530564[7]	AY326105[8]	GQ405391[9]	DQ023455[10]
Rattus rattus	NC_012374[11]	AM408328[12]	AM910976[13]	HQ334643[14]
Rattus tanezumi	EU273712[11]	DQ191515[15]	GQ405393[9]	KM397346[16]
Rattus norvegicus	AJ428514[17]	AJ429134[18]	NC_005101[19]	AY294938[20]
Rattus fuscipes	NC_014867[21]	HQ334623[14]	-	HQ334692[14]
Rattus leucopus	GU570659[21]	HQ334615[14]	EU349825[6]	EU349914[6]
Rattus niobe	KC152486[22]	HQ334580[14]	-	HQ334659[14]
Rattus praetor	NC_012461[11]	HQ334591[14]	GQ405392[10]	HQ334662[14]
Rattus lutreolus	GU570661[21]	HQ334613[14]	-	HQ334670[14]
Rattus sordidus	GU570665[21]	HQ334599[14]	-	HQ334691[14]
Rattus villosissimus	NC_014864[21]	HQ334576[14]	EU349826[6]	EU349915[6]
Rattus tunneyi	NC_014861[21]	HQ334579[14]	-	HQ334668[15]
Sundamys maxi				
RMNH 21479	**KY464170**	**KY467079**	**KY467090**	**KY467070**
RMNH 14208	**KY464171**	**KY467078**	**KY467089**	**KY467071**
Sundamys muelleri				
BOR448	**KY464172**	**KY467080**	**KY467091**	**KY467068**
EBD 30384M	**KY464173**	**KY467081**	**KY467092**	**KY467069**
Sundamys infraluteus				
BOR251	**KY464174**	**KY467083**	**KY467088**	**KY467073**
BOR510	**KY464175**	**KY467077**	**KY467087**	**KY467072**
*Rattus annandalei**				
MZB 28969	**KY464176**	**KY467082**	**KY467093**	**KY467074**
MZB 28971	**KY464177**	**KY467085**	**KY467086**	**KY467076**
Berylmys berdmorei	**KY464178**	HM217639[2]	-	-
Lenothrix canus	**KY464180**	**KY467084**	**KY467094**	**KY467075**
Niviventer confucianus	KJ152220	KC953416[5]	KC953293[5]	KC953540[5]
Niviventer excelsior	JQ927552[23]	DQ191511[15]	GQ405386[9]	-
Leopoldamys edwardsi	**KY464179**	HM217688[2]	-	KJ607312
Maxomys surifer	**KY464181**	HM217682[2]	DQ019065[10]	KM397347[16]
Micromys minutus	KP399599[24]	EU349862[6]	EU349818[6]	EU349904[6]

[1] Wang et al. 2015
[2] Pagès et al. 2010
[3] Achmadi et al. 2013
[4] Fabre et al. 2013
[5] Schenk et al. 2013
[6] Rowe et al. 2008
[7] Tsangaras et al. 2014
[8] Jansa and Weksler 2004
[9] Heaney et al. 2009
[10] Steppan et al. 2005
[11] Robins et al. 2008
[12] Michaux et al. 2007
[13] Lecompte et al. 2008
[14] Rowe et al. 2011
[15] Jansa et al. 2006
[16] Pisano et al. 2015
[17] Nilsson et al. 2003
[18] Huchon et al. 2002
[19] Rnor_6.0
[20] Steppan et al. 2004
[21] Robins et al. 2010
[22] McComish 2012
[23] Chen et al. 2012
[24] Jing 2015

*Emended to *Sundamys annandalei* in this study.

DNA extraction and sequencing

DNA was extracted using phenol-chloroform with ethanol precipitation or DNeasy Blood & Tissue Kit (Qiagen). Museum tissue samples from dried specimens were processed in an isolated ancient DNA laboratory. We used a modified Illumina protocol based on Maricic et al. (2010) to obtain complete mitogenomes from *Rattus annandalei*

MZB 28969, *Sundamys* species, and *Lenothrix canus* (Appendix 4.1). For *Berylmys berdmorei*, *Bunomys penitus*, *Halmaheramys bokimekot*, *Paruromys dominator*, *Rattus annandalei* MZB 28971, and *Maxomys surifer*, we obtained mitogenomes following the protocol of Tilak et al. (2015) and Fabre et al. (2016). These libraries were pooled and sequenced without enrichment as single-end reads on Illumina HiSeq 2000 lanes at the GATC-Biotech company (Konstanz, Germany). Nuclear genes were obtained following the protocol of Fabre et al. (2014, 2016) with some modifications (Appendix 4.1).

Genotyping and alignment of nuclear and mitochondrial sequences

We removed adaptors with cutadapt 1.8.3 (Martin 2011). Forward and reverse reads were paired in Geneious 8.1.5 (http://www.geneious.com, Kearse et al. 2012). We generated a *Sundamys* mitogenome reference by mapping reads from a modern *Sundamys muelleri* to *Rattus norvegicus* (AJ428514) with medium-low sensitivity and 5 iterations. We used this reference to map the rest of *Sundamys* samples, *Rattus annandalei* (MZB 28969) and *Lenothrix canus* in Geneious. For the nuclear genes, we mapped the reads to homologous sequences from *Rattus norvegicus* in Geneious using medium-low sensitivity and 3 iterations. We used SAMtools 0.1.18 (Li et al. 2009) to remove PCR duplicates from the mitochondrial and nuclear BAM mapping files and called consensus sequences in Geneious (parameters: minimum 2x and 75% threshold). For other libraries, raw 101 nucleotide (nt) reads were imported into Geneious, trimmed and iteratively mapped (minimum of 24 consecutive nt perfect match to reference, maximum 5% mismatch over read length, minimum of 3% of gaps with a maximum gap size of 3-nt) to the phylogenetically closest mitogenome available.

We used MAFFT 7.244 (Katoh et al. 2002) to align the mitogenomes. In the mitogenome alignments, we kept only protein-coding genes of the heavy strand (all genes except *nd6*) and excluded the rest. We also removed the overlapping region of *atp6* and *atp8* (43 nt), after confirming it evolves under a different evolutionary model than the sequences in the other protein-coding genes. This happens because in mitochondrial genomes the regions of adjacent genes can overlap, and thus can have stringent evolutionary constraints. The alignment was visually inspected and the genes were translated into amino-acids and inspected for stop codons in Geneious.

Sequences for each nuclear gene were aligned with MAFFT using the algorithm E-INS-i to overcome alignment problems caused by low homology between some of the

sequences, which could span slightly different regions in the same exon of the gene. The alignments were inspected visually and the genes were translated into amino acids and inspected for stop codons in Geneious.

We concatenated the mitochondrial and nuclear alignments into a supermatrix with the *ape* package in R (Paradis et al. 2004). This concatenated alignment was used for phylogenetic reconstructions and had 31 rows representing 27 species. For non-*Sundamys* species the concatenated sequences were chimeric, constructed from individuals sequenced in different studies (Table 4.1). Each row had 15,065 nt, of which 10,798 were mitochondrial and 4,267 were nuclear. The mitochondrial alignment had 0.07% missing data, and the nuclear alignment had 36% missing data.

Phylogenetic analysis and molecular dating

We used PhyloBayes 4.1 (Lartillot et al. 2009) for phylogenetic reconstruction based on the concatenated mitochondrial and nuclear supermatrix, as well as from each nuclear marker or mitochondrial DNA independently. PhyloBayes is a Bayesian Markov Chain Monte Carlo sampler, which incorporates methods for modeling site-specific sequence evolution variables from distributions not defined *a priori*, but inferred from the data. For each matrix, the CAT + GTR + Γ4 model was selected and 2 independent chains were run for 10,000 cycles and sampled every 10 generations with a burn-in of 1,000 trees. All runs showed good convergence since the maximum differences of the bipartition frequency between both chains was < 0.1. We used a sequence from *Micromys minutus* to root the trees.

For comparative purposes we computed pairwise genetic distances (proportion of nucleotides at which 2 sequences differ) among recognized species in *Sundamys*, *R. annandalei,* and the well-studied *R. exulans* in MEGA 6.06 (Tamura et al. 2013). Distances were calculated separately for the 3 concatenated nuclear markers and for mitogenomes.

We inferred evolutionary relationships and dated divergences in a Bayesian framework with BEAST 2.4.4 (Bouckaert et al. 2014). For this purpose, we used a mitogenome matrix with only protein-coding genes and one sequence per species to meet the tree prior of the model. To include a calibration point, we incorporated mitogenomes from 6 murines from 2 molecular tribes from the "*Mus*" branch of the *Rattus/Mus* split (Apodemini: *Apodemus chejuensis* HM034867, *A. latronum* NC_019585, *A. peninsulae*

NC_016060; Murini: *Mus cervicolor* KJ530560, *M. cookii* KJ530561, *M. spretus* NC_025952) (Fabre et al. 2013; Pagès et al. 2015). We determined the best partition scheme with PartitionFinder 2.1.1 (Lanfear et al. 2016). It split the alignment into 3 sets corresponding to the 3 codon positions, all of which evolved under a GTR + Γ + I model, except codon position 3 of *nd6* which fell in its own partition and was discarded for downstream analysis. We split the mitogenome alignment into the former 3 partitions with AMAS (Borowiec 2016) and imported them into BEAUTi. We assigned an independent GTR + Γ + I model with estimated base frequencies and estimated substitution rate for each partition. We linked for the 3 partitions an uncorrelated relaxed clock model with rates sampled from a lognormal distribution and set a Yule model of speciation process as a tree prior. We used 11.81 My (95% CI: 11.11-12.68 My) as a prior for the split between the in-group Rattini and the incorporated *Mus*-related lineages, as suggested in Kimura et al. (2015). This prior was specified in BEAST as a lognormal distribution as suggested in Morrison (2008). The *Mus*/*Rattus* split interval in Kimura et al. (2015) is based on a well-represented phylogeny of Murinae with nuclear and mitochondrial DNA (Fabre et al. 2013) with an extra calibration point from a new fossil of the *Mus/Arvicanthis* split. It matches the 11.0-12.3 My interval reviewed in Benton and Donoghue (2007) from *Progonomys* and *Karnimata* fossils, which has already been used for dating divergences in a mitogenome phylogeny of *Rattus* (Robins et al. 2008). Aplin et al. (2011), however, proposes a more relaxed interval (i.e. 10.4-14 My) should be used to incorporate the uncertainties surrounding the fossil record of this group. We ran 2 chains in BEAST 2.4.4 on the XSEDE cluster via the Cipres Science Gateway (Miller et al. 2010) for 50 million generations, sampled every 10,000. We assessed the convergence between the 2 chains in Tracer by confirming the estimated sample size (ESS) was > 200 for each of the parameters in the combined log file, after discarding the first 10% of generations. We discarded the first 10% of the trees from each chain and combined them to form the posterior. A maximum clade credibility tree was generated with TreeAnnotator.

Geometric morphometric procedures

Photographs were taken of 122 Sundamys and Rattus annandalei specimens, as well as for 83 other Rattus specimens belonging to 9 species (*R. andamanensis, R. argentiventer, R. baluensis, R. exulans, R. losea, R. norvegicus, R. rattus, R. tanezumi,* and *R. tiomanicus*; Appendix 4.2 and 4.3). We targeted a sample size of 30 adult

individuals per species for Sundamys and 10 individuals for the Rattus species with an equal number of males and females. The specimens studied here are stored at the American Museum of Natural History, New York, USA (AMNH); Natural History Museum, London, UK (BMNH); Delaware Museum of Natural History, Wilmington, Delaware (DMNH); Muséum National d'Histoire Naturelle, Paris, France (MNHN); Centre de Biologie pour la Gestion des Populations, Montpellier, France (CBGP); Estación Biológica de Doñana, Seville, Spain (EBD); Field Museum of Natural History, Chicago, Illinois (FMNH); Museum of Comparative Zoology, Harvard University, Cambridge, Massachusetts (MCZ); the Museum Zoologicum Bogoriense, Cibinong, Indonesia (MZB); Naturalis Biodiversity Center, Leiden, Netherlands (RMNH); National Museum of Natural History, Smithsonian Institution, Washington, D.C., USA (USNM); and Zoological Museum – University of Copenhagen (ZMUC). We carefully checked the skin and skull to avoid any misidentifications. All Rattus specimens from the CBGP were also molecularly identified (Pagès et al. 2010, 2013). Dental wear patterns were used to avoid photographing juveniles. To explore morphological variation, 25 landmarks were placed on the palatal view of the cranium for 205 specimens and 18 landmarks on the lateral view of the dentary for a subset of 95 specimens (Figure 4.2). A CANON 7D video camera equipped with a macro-lens EF 100mm f/2.8L and the software TPS dig2 (Rohlf 2013) was used to obtain the photographs. We tested repeatability with 30 repetitions of land mark placement on 3 specimens of Rattus exulans. Measurement error was evaluated with a Procrustes ANOVA as in Claude (2013). The among- and within-specimen variances were computed based on the mean squares and cross products corresponding to the specimen and residual sources of variation. The percentage of measurement error was less than 1% for both dentary and palatal centroid size, and 8% and 6% for the palatal and dentary shape, respectively.

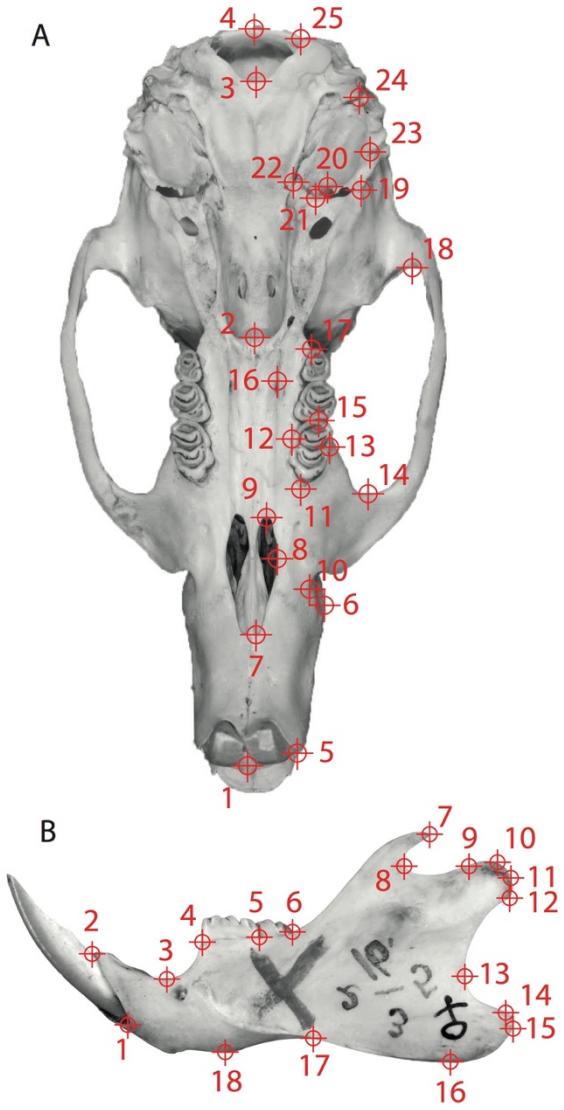


Figure 4-2. Landmark locations and definitions on the palatine and dentary views of the *Sundamys* and *Rattus* specimens. A) palatal side: 1) premaxillary bone between anterior margin of upper incisors, 2) posterior margin of the palatal bone, 3) anterior margin of the foramen magnum, 4) posterior margin of the foramen magnum, 5) lateral margin of the right incisor, 6) lateral marginal of the lacrymal notch, 7) anterior margin of the incisive foramina, 8) lateral margin of the incisive foramina, 9) posterior margin of the incisive foramina, 10) maxillary insertion of the zygomatic root, 11) anterior margin of the first upper molar, 12) lingual margin of M1 at the level of the second lamina, 13) labial margin of M1 at the level of the second lamina, 14) anterior margin of the orbit, 15) posterior margin of M1, 16) posterior margin of the posterior palatine foramina, 17) postero-lateral margin of M3, 18) posterior margin of the temporal fossa, 19) suture between squamosal and sphenoid, 20) junction between bulla tympanica and eustachian tube, 21) lateral tip of the eustachian tube, 22) junction between basioccipital and basisphenoid, 23) most internal point of the external auditory meatus, 24) contact between bulla tympanica and jugular process, 25) lateral margin of the foramen magnum. B) Dentary lateral side: 1) ventral margin of mandible at insertion of incisor, 2) dorsal margin of mandible at insertion of incisor, 3) lowest point on dorsal surface of mandible between incisor and m1, 4) anterior margin of m1 at insertion, 5) insertion of m2 at posterior margin, 6) intersection of the ascending branch of the mandible and the molar row, 7) dorsal-most point of the coronoid process, 8) deepest point of condylar neck, 9) anterior-most point of the articular condyle, 10) dorsal-most point of the articular condyle, 11) posterior-most point of the articular condyle, 12) ventral-most of point the articular condyle, 13) anterior-most point between articular condyle and angular process, 14) dorsal-most point of the angular process, 15) posterior-most point of the angular process, 16) ventral-most point of the angular process, 17) dorsal-most point on the ventral margin of mandible anterior to angular process, 18) ventral-most part of the mandible anterior to 17.

We used classic geometric morphometric methods (Bookstein 1991; Slice 2007; Adams and Otárola-Castillo 2013) to provide a description of the shape of the palatal view of the skull and dentary as well as to locate the most variable parts of the skull and dentary among Sundamys and Rattus species, and within Sundamys species. Landmark coordinates were analyzed using a general Procrustes analysis (GPA, Rohlf and Slice 1990). The logarithm of the centroid size was used as an indicator of size. A principal component analysis (PCA) was computed on superimposed coordinates (Dryden and Mardia 1998) and the scores of the principal components (PCs) were used in the multivariate analyses. We computed extreme morphologies along the first 2 PC axes to

visualize patterns of shape variation. A 3-way linear discriminant analysis (LDA) was also computed on Rattus and Sundamys genera factors, with R. annandalei as an unknown factor. Thus, 3 factors were set in the LDA: Rattus, Sundamys, and R. annandalei. We subsequently computed the predicted values for R. annandalei following the protocol of Claude (2013). A multivariate analysis of covariance (MANCOVA) was run using centroid size as a co-variate to test the effects of species and sex. A multivariate linear model was also applied to the PCs of shape variation using all axes with non-null eigenvalues to see the potential effect of 4 explanatory variables. The explanatory variables considered here were the minimum and maximum elevational ranges, species, size, and interactions until the third order between the variables. Full factorial MANOVA was used to test the effects of species, size, sex, and elevation (lowland versus highland) on skull and dentary shapes.

Results

Sequence data

Complete mitochondrial genomes were sequenced for 15 murines from 11 species, GenBank KY464167 - KY464181. The nuclear loci (exon 10 of *ghr*, partial exon 1 of *rbp3*, and partial exon 1 of *rag1*) were sequenced for 9 animals from 5 species, GenBank KY467068 - KY467094. The coverage for mitochondrial genomes ranged from 5x to 192x. Sequences for the other taxa were downloaded from GenBank (Table 4.1) to complete a final supermatrix with 31 samples from 27 species.

Phylogenetic results and molecular dating

The phylogeny inferred from the mito-nuclear supermatrix was highly resolved, with a posterior probability (*PP*) of 1.00 for most nodes in the tree, including the clade for *Sundamys* and its internal species relationships (Figure 4.3). *Rattus annandalei* is nested within *Sundamys* as the sister to *S. infraluteus* (*PP* = 1). The tree supports a sister relationship between *S. muelleri* and *S. maxi* (*PP* = 1). The genus *Berylmys* appears as sister to all other rats in the Rattus division included in this phylogeny. The rats from the Dacnomys division, *Leopoldamys*, *Niviventer*, and *Lenothrix*, form a well-supported clade sister to the Rattus division. A tree based on mitochondrial DNA alone was very similar to the concatenated mito-nuclear tree. However, the trees based on individual

nuclear loci were not sufficiently resolved to recover well supported relationships between *Sundamys* and other genera, or within *Sundamys* (Appendix 4.4).

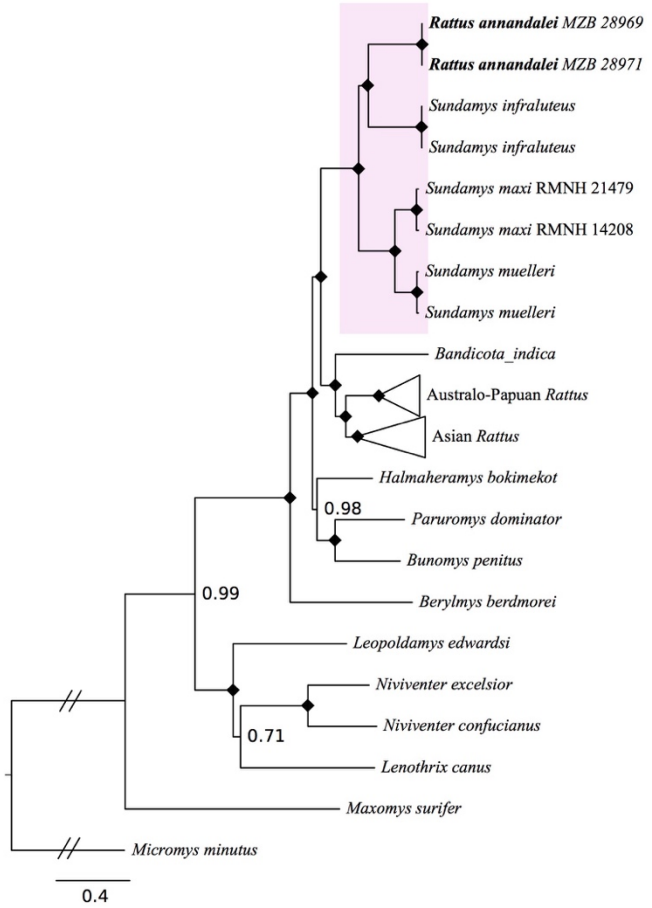


Figure 4-3. PhyloBayes maximum clade credibility tree for the concatenated mitochondrial protein-coding genes on the heavy strand (10,798 bp) and nuclear loci (*rbp3*, *ghr*, and *rag1*, 4267 bp). Posterior probabilities (*PP*) are indicated by diamonds when *PP* = 1.00, otherwise with a number. Scale bar represents substitutions per site.

The time to the most recent common ancestor (TMRCA) of *Sundamys* and other Rattini was estimated at 3.88 million years ago (My) (95% HPD: 3.26–4.52 My, Figure 4.4). At 2.69 My (2.16–3.18 My) *Sundamys* split into the ancestors of *S. muelleri* + *S. maxi* and *R. annandalei* + *S. infraluteus*. The split between *S. infraluteus* and *R. annandalei* was estimated at 2.22 My (1.74–2.71 My), whereas the split between *S. maxi* and *S. muelleri* was more recent: 1.22 My (0.90-1.57 My).

The mitochondrial and nuclear distances between *R. annandalei* and any species of *Sundamys* were (1) smaller than between it and a supposed *Rattus* congener, and (2) within the range of distances among recognized *Sundamys* species (Table 4.2). Mitochondrial distances were approximately an order of magnitude greater than nuclear distances.

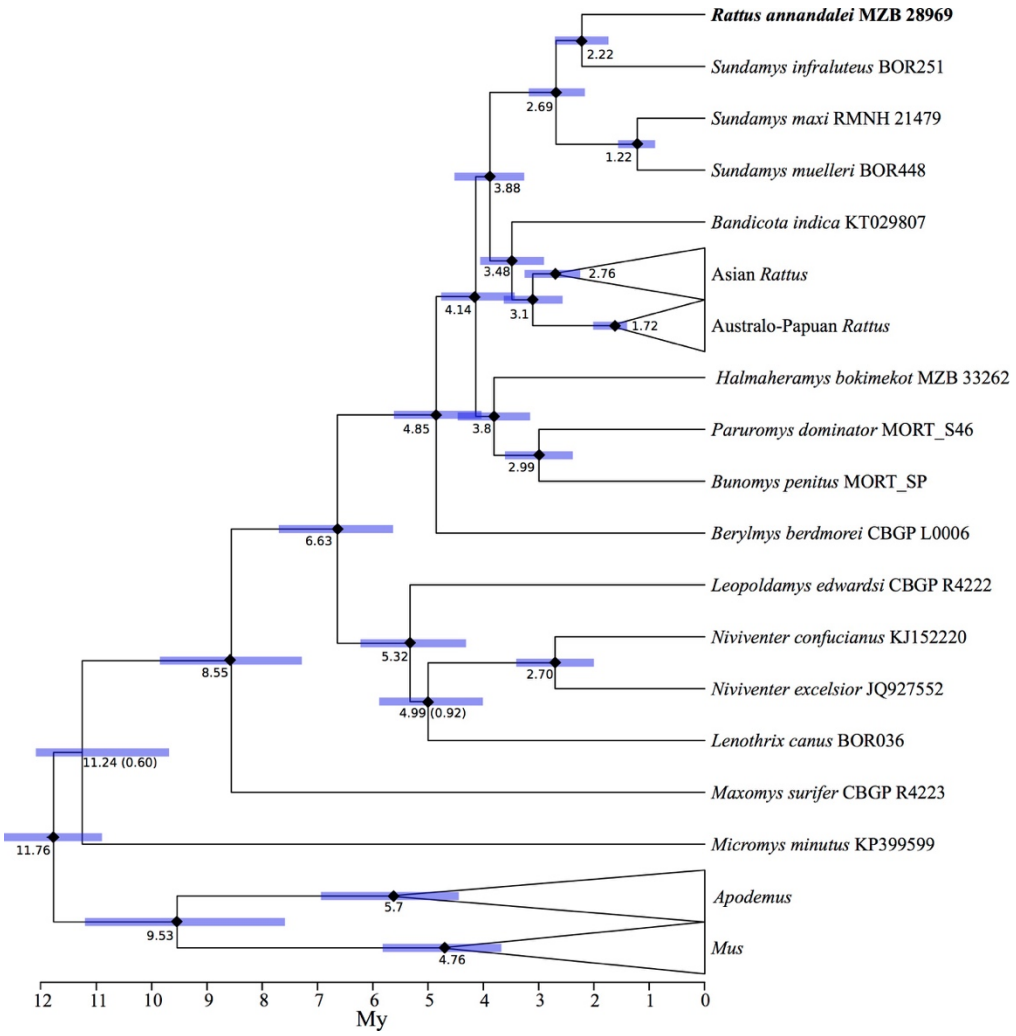


Figure 4-4. Maximum clade credibility tree from BEAST analysis of the protein-coding genes on the heavy strand of the mitochondria. Node bars indicate 95% highest posterior density in divergence times. Numbers on nodes represent ages in My. Posterior probabilities equal to 1.00 are indicated with diamonds on the nodes, or otherwise with a number in parentheses.

Table 4-2. Uncorrected pairwise genetic distances (*p*) between *R. annandalei*, all *Sundamys* species and a *Rattus* species with a well-defined taxonomy, *Rattus exulans*, for mitogenomes (upper triangular) and concatenated nuclear markers *rbp3*, *rag1*, and *ghr* (lower triangular).

	1	2	3	4	5
1. *Rattus annandalei*		0.095	0.102	0.101	0.120
2. *Sundamys infraluteus*	0.012		0.101	0.098	0.116
3. *Sundamys maxi*	0.012	0.013		0.062	0.119
4. *Sundamys muelleri*	0.010	0.014	0.005		0.118
5. *Rattus exulans*	0.022	0.021	0.025	0.023	

Morphometric results from Sundamys–Rattus comparison

Palatal analyses of *Rattus* and *Sundamys* species clearly illustrate the mixed features of *R. annandalei* (Figure 4.5 A). PC1 and PC2 respectively explained 45.3% and 8.3% of

the variance. PC1 separated *Rattus* from *Sundamys*. *Rattus* species tend to have a shorter rostrum, longer incisive foramina, longer palatal bridge, wider bullae tympanica, narrower incisors, shorter molar rows, a wider braincase, and a longer basicranium (Figure 4.5 A, see variation on PC1, black line). *Sundamys* species tend to have a longer rostrum, shorter bullae tympanica with a well-defined Eustachian tube, wide incisors, a narrow skull, and a shorter braincase (Figure 4.5 A). The morphospace of *R. annandalei* falls between those of *Rattus* and *Sundamys*, a pattern fully detailed in the following emended diagnosis. A 3-way linear discriminant analysis was subsequently computed to estimate the predicted morphological attribution of *R. annandalei*. Once again *R. annandalei* fell in a singular morphospace with intermediate values between *Sundamys* and *Rattus* for the LD1 axis (Leave-one out cross-validation on LD1, *CV1*, = 31.9%), that best discriminates *Rattus* and *Sundamys,* and LD2 (*CV2* = 12.6%) which isolates *R. annandalei* from the other discriminant factors (*Rattus* and *Sundamys*) (Figure 4.5 B). Taking *Rattus annandalei* as an unknown factor, 14 specimens were attributed to *Rattus* and 19 specimens to *Sundamys*.

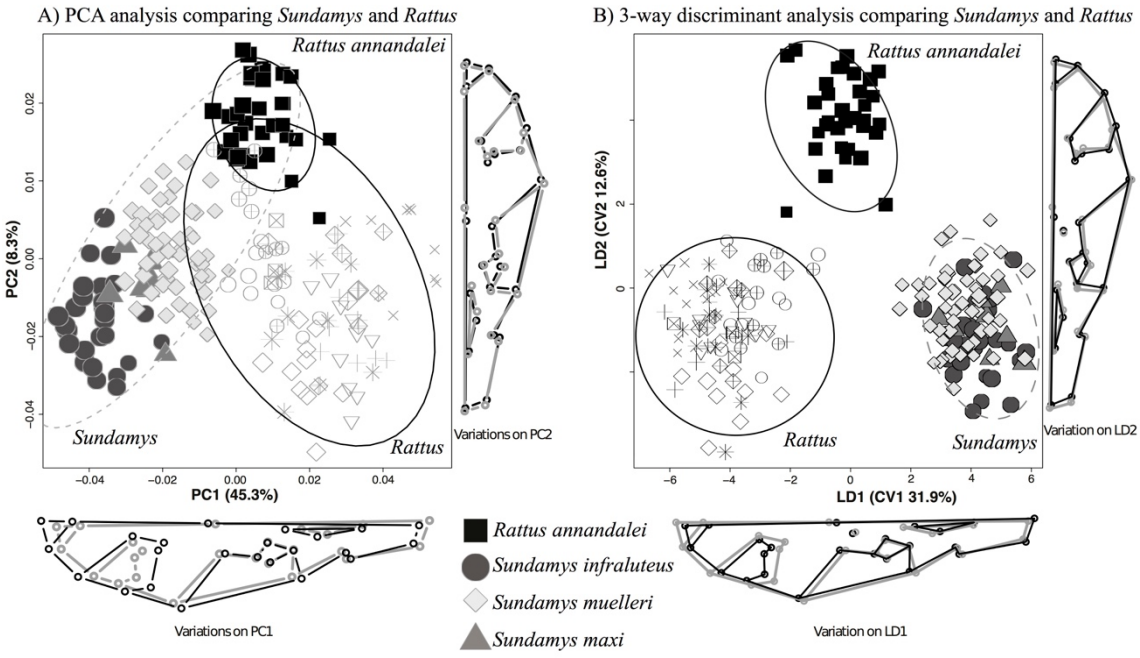


Figure 4-5. A) Principal components and B) 3-way discriminant analyses of morphological variation for the cranium (palatal view) among *Sundamys* and Indo-Pacific *Rattus* species. Patterns of shape variation along PC1 and PC2 (A) and LD1 and LD2 (B) are illustrated on the right and below each graph. Light grey lines and circles correspond to minimal scores, and black lines and circles to maximal values. Symbols are proportional to centroid skull and dentary size. Open symbols correspond to *Rattus* species and closed symbols to *Sundamys*. Ellipses show 95% confidence area for each genus and *R. annandalei*. List of *Rattus* and *Sundamys* species as well as their voucher numbers are indicated in Appendix 4.2.

Morphometrics of the Sundamys Cranium

PC1 and PC2 respectively explained 38.4% and 9.0% of the variance (Figure 4.6 A). The first principal component discriminates *Rattus annandalei,* in the negative region of PC1 and *S. infraluteus*, in the positive region*,* whereas *S. maxi* and *S. muelleri* occupy central positions, with *S. maxi* closer to *S. infraluteus*. This axis is correlated with smaller braincase, smaller tympanic bullae with a small Eustachian tube, and a small first upper molar (Figure 4.6 A). A MANOVA computed on the PC scores revealed a highly significant effect of species ($F = 30.9$; $d.f. = 3$; $P < 0.0001$), elevation ($F = 128.2$; $d.f. = 1$; $P < 0.0001$), and size ($F = 39.8$; $d.f. = 1$; $P < 0.0001$), but no effect of sex ($F = 1.5$; $d.f. = 2$; $P = 0.07$). No significant interaction ($P = 0.20$) was detected between species and size ($F = 1.28$; $d.f. = 2$; $P < 0.02$) or between elevation and size ($F = 1.4$; $d.f. = 11$; $P = 0.19$). A MANCOVA performed on centroid size indicated significant effects of species ($F = 17.4$; $d.f. = 2$; $P < 0.0001$) and elevation ($F = 38.6$; $d.f. = 1$; $P < 0.0001$).

Morphometrics of the Sundamys Mandible

PC1 and PC2 represented 32.3% and 16.1% of the explained variance, respectively (Figure 4.6 B). These axes are less discriminating compared to the palatal view of the skull. PC1 separated the dentary of lowland species (*Rattus annandalei* and *S. muelleri*) and highland species (*Sundamys infraluteus* and *S. maxi*) (loadings in Supplementary Data S8). The dentary of the highland species is more dorso-ventrally compact. In highland *Sundamys*, the angular process is shorter and does not extend ventro-posteriorly to the articular process. The m1 as well as the molar row is also longer in these highland species. PC2 is mainly correlated with the age of the specimen and size. The coronoid process of the highland species is shorter and situated closer to the large condyloid process (Figures 4.6 B and 4.7). A MANOVA computed on the PC scores revealed highly significant effects of species ($F = 8.70$; $d.f. = 3$; $P < 0.0001$) and size ($F = 11.4$; $d.f. = 1$; $P < 0.0001$), non-significant interaction between species and size ($F = 1.3$; $d.f. = 3$; $P = 0.22$), and non-significant effect of sex ($F = 0.61$; $d.f. = 2$; $P = 0.76$). A MANCOVA computed on centroid size show highly significant effects of species ($F = 22.99$; $d.f. = 2$; $P < 0.0001$) and elevation ($F = 21.99$; $d.f. = 1$; $P < 0.0001$).

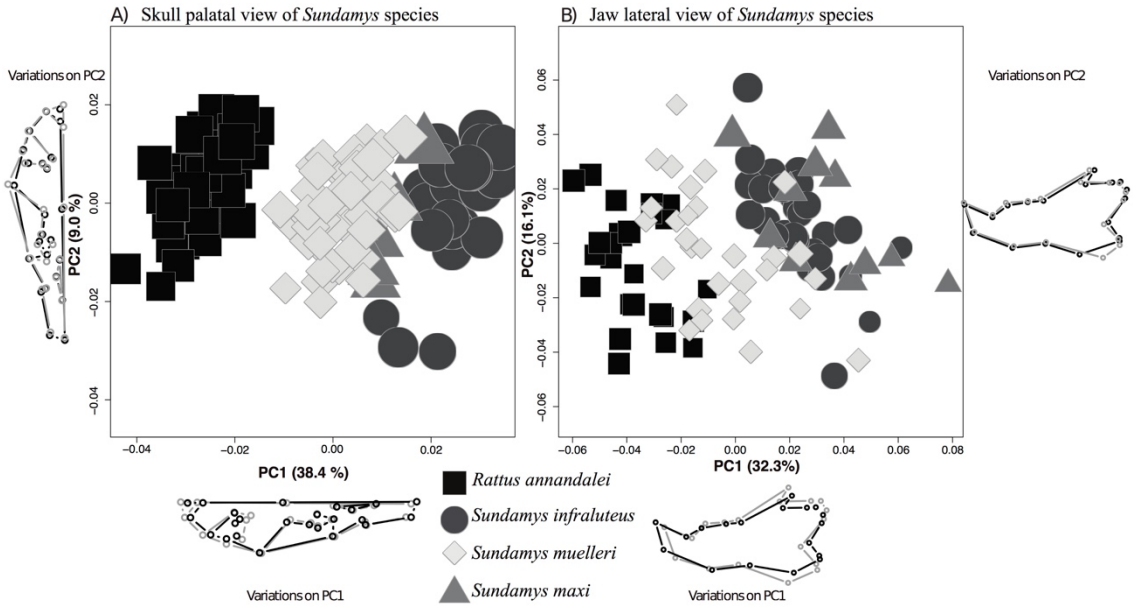


Figure 4-6. Principal components analysis for the palatal view of the cranium (A) and lateral view of the mandible (B) among *Sundamys* species and *Rattus annandalei*. Patterns of shape variation along PC1 and PC2 are illustrated on the side and below each graph, with light grey lines and circles corresponding to minimal scores, black lines and circles corresponding to the maximal scores. Symbols are proportional to centroid skull and dentary size. List of *Sundamys* species as well as their voucher numbers are indicated in Appendices 4.2 and 4.3.

Emended diagnosis

Genus *Sundamys* Musser and Newcomb (1983)

Type species.— *Mus mülleri* Jentink (1879).

Phylogeny.— *Sundamys* belongs to the Rattus division of the Rattini tribe in the Murinae subfamily. It is closely related to the widespread genus *Rattus*; the Philippine genus *Bullimus*; Sulawesi genera *Bunomys*, *Paruromys*, and *Taeromys*; and the Moluccan genus *Halmaheramys* (see also Fabre et al. 2013).

Diagnosis.— Once considered part of *Rattus*, *Sundamys* was defined by Musser and Newcomb (1983) to include 3 species: *Sundamys muelleri*, *S. infraluteus*, and *S. maxi*. Our molecular phylogenies clearly support the inclusion of *Rattus annandalei* within *Sundamys*. Based on this new result, we provide an emended diagnosis of *Sundamys*. *Sundamys* have (1) medium to large body size (body-mass range: 150-643 grams; Table 4.3) compared to other Rattini from the Sundaic region (e.g. *Niviventer*, *Maxomys*); (2) a slightly inflated rostrum without a marked constriction anterior to the lacrymal notch;

(3) 4 pairs of mammae (1 pectoral + 1 post-axillary + 2 inguinal) are found in all species except *S. infraluteus* (1 post-axillary + 2 inguinal); (4) incisive foramina are short relative to condylobasal length and usually do not reach anterior margin of M1 (Figure 4.7); (5) small sphenopalatine vacuity; (6) upper molars are anchored by 5 (M1), 4 (M2), or 3 roots (M3); (7) lower molars are anchored by 4 (m1) or 3 roots (m2, m3); (8) palatal bridge extends only slightly beyond the molar rows without forming a wide and deep shelf (Figure 4.7); (9) mesopterygoid fossa is wide and connects with the sphenopalatine vacuity; (10) the posterior cingulum is often present on M1; (11) M2 and M3 usually have large and well-developed cusp t3 (sensu Musser 1981; Musser and Newcomb 1983); (12) lower molars have wide lamina made of similar size cuspids that do not form lamina with arcuate or acute angles; (13) anterolabial and anterolingual cusps of m1 are fused and form a large lamina; (14) the antero- and postero-labial cusplets on m2 are always present and well-developed.

Content and distribution.— *Sundamys* includes 4 species confined to the Sunda Shelf region: *Sundamys annandalei*, *S. infraluteus*, *S. maxi*, and *S. muelleri*.

Table 4-3. Selected external measurements (mm) and body mass (g) of adult *Sundamys* species from the main landmasses. Mean, sample size (in parenthesis) and range (in brackets) are reported in each case.

	Sundamys muelleri			*S. infraluteus*		*S. maxi*	*S. annandalei*	
	Malay Peninsula	Sumatra	Sabah	Sabah	Sumatra	Java	Malay Peninsula	Sumatra
Head-Body	243.1 (58) [209-299][1]	207.3 (23) [185-236][1]	207.7 (22) [188-240][1]	258.5 (10) [229-282][1]	266 (3) [259-276][1]	241.5 (8) [218-270][1]	191.8 (24) [173-220][1]	192
Tail	285.6 (58) [248-370][1]	260.6 (22) [214-301][1]	247.4 (22) [212-271][1]	315.8 (10) [289-343][1]	311.7 (3) [298-333][1]	286.7 (7) [258-309][1]	240.2 (24) [225-263][1]	227
Hindfoot	51.5 (62) [47-55][1]	45.3 (23) [42-49][1]	40.9 (22) [37-45][1]	57.5 (11) [55-61][1]	58.7 (3) [57-60][1]	53.1 (7) [52-55][1]	39.5 (24) [37-41][1]	38.94
Ear	23.2 (57) [20-27][1]	21.4 (23) [20-23][1]	21.4 (7) [21-23][3]	24.8 (10) [22-27][1]	25.7 (3) [24-29][1]	25.4 (7) [24-28][1]	21.2 (24) [20-23] [1]	21.31
TL / HB	116[1]	121[1]	119[1]	122[1]	117[1]	119[1]	125	118
Body mass	335.4 (30 ♂) 292.4 (30 ♀)[2]	NA	262 (8 ♂) 196 (3 ♀)[3]	468 (2 ♂) 550 (2 ♀)[3]	582 (2) [521-643] [1]	NA	197.4 (15) [155-261] [1,4]	150

[1] Musser and Newcomb 1983, Pp: 428, 442, 452, 502 (claws included in hindfoot measurement).

[2] Boo-Liat 1970

[3] Field measurements.

[4] MNHN and USNM.

Species *Sundamys annandalei* (Bonhote, 1903).

Type locality: Malaysia (Malay Peninsula), South Perak, Sungkei.

Emended comparison with other Sundamys species.— *Sundamys annandalei* is smaller (head-body range: 173–220 mm; body mass range: 150-261 grams) than other *Sundamys* (head-body range: 185–299 mm; body-mass range: 196–643 grams; Table 4.3). Its dorsum is grayish-brown and its belly ranges from white to pale yellow, grayish white or buff white. It is very similar in color pattern to *Sundamys muelleri,* although the taxa differ in body size and proportions. Indeed, *S. muelleri* is a larger species and can be twice the weight of *S. annandalei*. The fur of *S. annandalei* is shaggy, and soft, with some longer guard hairs present on the rump. Compared to *S. annandalei*, both *S. maxi* and *S. infraluteus* are giant rats with dull, soft, dark fur on the dorsum and paler on the belly. Tail length is longer than head-body length in all *Sundamys* species. One major difference among *Sundamys* species is the foot length, which is the shortest in *S. annandalei* (37-41 mm) and largest in *S. infraluteus* (55-61 mm) (Table 4.3). *Sundamys annandalei* has the following mammae formula: 1 pectoral + 1 post-axillary + 2 inguinal, with a total of 8 teats.

Similar to overall body size, the skulls of *S. maxi* and *S. infraluteus* are much larger than those of *S. annandalei* and *S. muelleri* (Figure 4.7). Dorsally the skull of all *Sundamys* are very similar in proportion apart from the ridging of the supra-orbital and temporal regions, which is more marked in the larger *S. infraluteus* and *S. maxi*, likely due to larger temporalis muscle in these species. With the exception of *S. maxi*, all *Sundamys* species have the common murid arterial pattern (following Musser and Newcomb 1983) with the stapedial artery branching into the tympanic bulla and branching laterally to become the internal maxillary artery. In *Sundamys maxi*, this branch is reduced or absent, and the internal maxillary artery is branching from the main internal carotid artery. The position of the transverse canal is generally anterior to the posterior opening of the alisphenoid canal.

In palatal view, most *Sundamys* species have the zygomatic plate of the zygomatic arch placed anterior to M1. However, in *S. infraluteus* the zygomatic plate overlaps with M1 along the antero-posterior axis. In *S. annandalei* the squamosal root of the zygomatic arch is near the tympanic bullae. Other *Sundamys* species have reduced tympanic bullae not overlapping with the squamosal root of the zygomatic arch along the antero-posterior axis. In palatal and lateral view, the major distinction of *S. annandalei* compared to other *Sundamys* species is its large and inflated tympanic

bullae. The tympanic bulla is smaller in all other species of *Sundamys*. Except for some individuals of *S. maxi*, an alisphenoid strut is present in most specimens. In *S. annandalei* and *S. maxi*, the sphenopterygoid vacuity is present, being much larger in most specimens of *S. maxi* (see Musser and Newcomb 1983, p 467). In the other species this vacuity is closed with a bony wall. Except for some *S. infraluteus* specimens, the sphenoid and vomer bridge is always present and well visible between the mesopterygoid fossa.

Concerning molar teeth, most *Sundamys* species have 5 roots under the M1, 4 roots under M2, and 3 roots under M3. It is only in *S. muelleri* that some specimens were found to have either 3 or 4 roots on M3. On the lower molars we observed no variation, with 4 roots under m1 and 3 roots under m2 and m3. Compared to the skull length, the molars are small in both *S. annandalei* and *S. muelleri*. Another state is found in *S. infraluteus* and *S. maxi* where both species have large molars relative to skull length. However, despite this large molar size, the cusp pattern of *S. maxi* is more similar to that of *S. muelleri* (Musser and Newcomb 1983). On the M1, a posterior cingulum is present in all *Sundamys* species. On M2 and M3, the cusp t3 is present and well developed in most of the observed specimens of *S. muelleri, S. maxi*, and *S. annandalei*. Concerning the lower cheek teeth, *S. annandalei* and all other species of *Sundamys* lack a posterior cingulum on m3 and they all have a postero-labial cusplet on m1. Antero-labial cusplet on m1 is often absent (in 70% of the observed specimens). Antero- and postero-labial cusplets are usually present on m2 in all *Sundamys* species. The same is true for the postero-labial cusplets on m3.

The dentary distinguishes high elevation *Sundamys* species from the low elevation ones. As discussed in our morphometric results, both lowland *S. muelleri* and *S. annandalei* have proportionally less elongated dentary compared to highland *S. infraluteus* and *S. maxi* ones (Figure 4.6 B). The angular process of *S. annandalei* and *S. muelleri* is wider and more posterior-ventrally elongated as compared to *S. infraluteus* and *S. maxi*, only slightly extending posteriorly beyond the condyloid process (Figure 4.6 B). Another clear-cut difference is the length of the lower molar row, which is longer in the highland *S. infraluteus* and *S. maxi*. The rami are also shorter in the highland *S. infraluteus* and *S. maxi*. The higher coronoid along with robust dentary (with a short and wide angular) constitute a more powerful in-lever arm for gnawing with high mechanical advantage for these lowland species. Further, the wide and robust

angular configuration indicates a larger insertion for the superficial masseter, which reinforces their dentary lever advantage at the incisor.

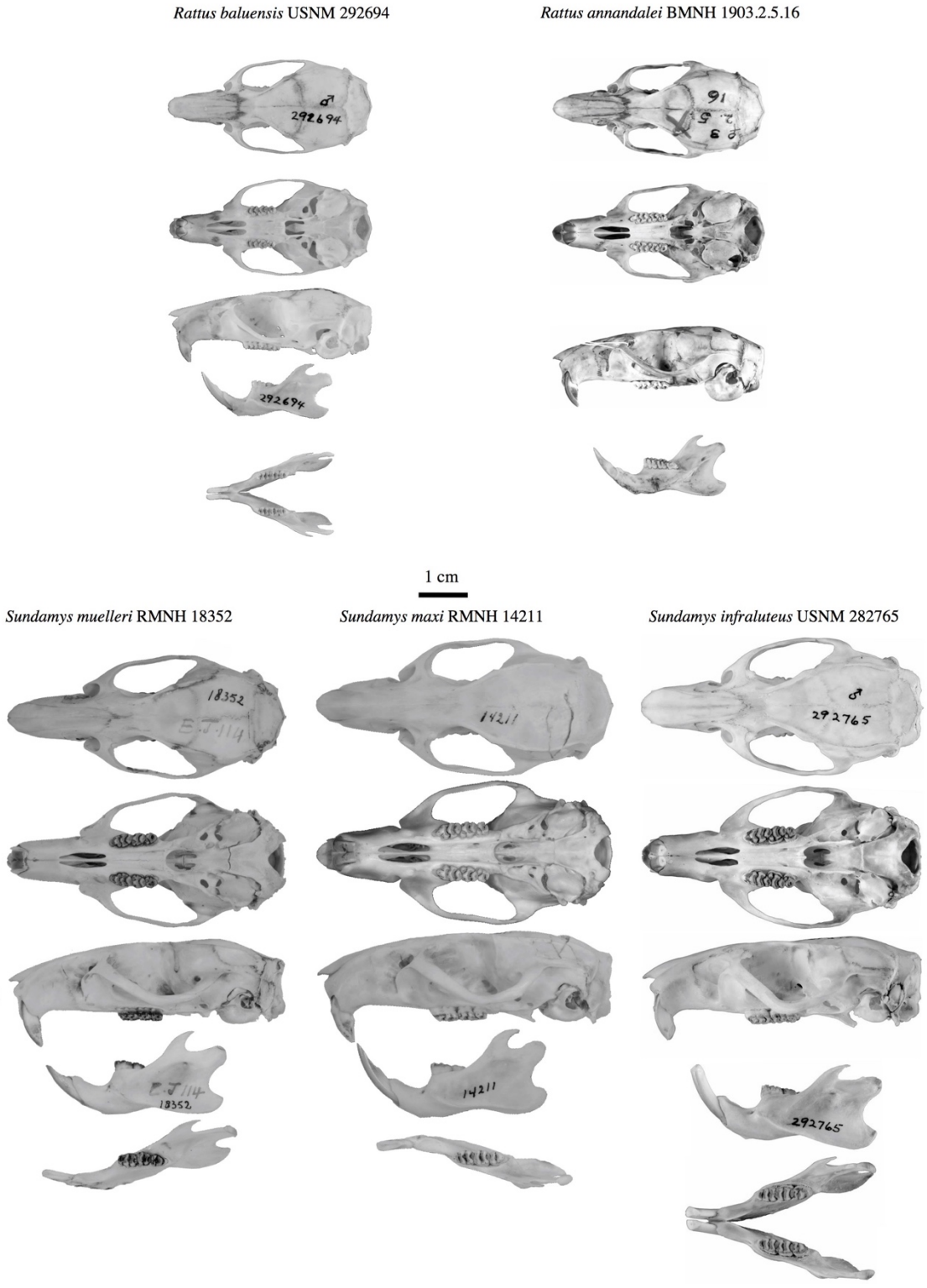


Figure 4-7. Dorsal, palatal, and lateral views of the skull, plus lateral and occlusal views of the dentary of *Rattus baluensis*, *Sundamys annandalei*, and the 3 previously recognized *Sundamys* species (*S. muelleri*, *S. maxi*, and *S. infraluteus*).

Discussion

Taxonomic status of Sundamys annandalei

Based on our molecular phylogenies inferred from nuclear and mitochondrial DNA and the morphological data, we reclassify *Rattus annandalei* as *Sundamys annandalei*. This taxonomic change is supported by the previous morphological revision of this species (Musser and Newcomb 1983) and our emended diagnosis of *Sundamys*. It resolves a longstanding debate on the taxonomic status of this taxon. Symplesiomorphies in the skull and dentition caused the confusion surrounding *S. annandalei* and also hinder the taxonomy of other Indo-Australian murids (Musser and Newcomb 1983). Several species will likely be revised in the near future (see also Carleton and Musser 2005), such as the extinct *Rattus macleari* and *R. nativitatis* from the Christmas Islands, or the Flores and Timor rats (*R. timorensis* and *R. hainaldi*) (Musser and Newcomb 1983). Museums are a key resource for better understanding the evolution and diversity of the Rattini, since some of the lineages are already extinct (*R. macleari, R. nativitatis*), have an unknown conservation status (*S. maxi*), or have only been recorded by a handful of specimens in museums (e.g. *R. blangorum, R. korinchi*, and *Pithecheirops otion*). Further molecular studies are required to test for the monophyly of *Rattus* and to more accurately define taxonomic groups within the Indo-Pacific region (Rowe et al. 2011; Fabre et al. 2013).

Sundamys biogeography and ecology

Our nuclear and mitochondrial analyses indicate close affinities between *Sundamys* and the genera *Rattus*, *Paruromys*, *Bunomys*, *Halmaheramys,* and *Bandicota*, as also identified in the most recent studies of these groups (Fabre et al. 2013; Schenk et al. 2013). Assuming our date estimates are reasonably accurate, the *Sundamys* lineage diverged from other Rattini in the Pliocene, ~2.69-3.88 My. Other studies on Indo-Pacific murids find that many of the genera in Rattini originated during the late Pliocene, following range expansions to new archipelagos (i.e. Sahul, Philippines, Wallacea, and Sundaland), whereas most of the intra-generic diversification has occurred within archipelagos during Late Pliocene and Pleistocene (Fabre et al. 2013; Schenk et al. 2013). This is coherent with a dynamic mosaic of intermittent physical (sea-level) and ecological (drier vegetation during glacial periods) barriers during the Plio-Pleistocene, which seem to have shaped much of the diversification of forest

mammals in Sundaland (Ruedi and Fumagalli 1996; Esselstyn et al. 2013; Leonard et al. 2015; Demos et al. 2016). A Pleistocene origin is suggested for *S. infraluteus* (~2.22 My) and other Sunda mountain small mammals, such as *Rattus baluensis* (Aplin et al. 2011) or mountain shrews in Java and Sumatra (Esselstyn et al. 2013; Demos et al. 2016), but contrasts with the pre-Pleistocene origin of highland lineages in Sunda squirrels (den Tex et al. 2010; Hawkins et al. 2016) and treeshrews (Roberts et al. 2011). Whereas all of these date estimates are crude, we need more nuclear markers and a wider sampling across the Sunda shelf to assess the effects of Late Pliocene - Pleistocene changes within and between species of *Sundamys*.

Sundamys muelleri and *S. annandalei* are sympatric across the entire distribution of *S. annandalei* (eastern Sumatra and southern Peninsular Malaysia, Figure 4.1). These 2 species have been trapped in the same surveys (Rudd 1965; Shariff 1990), but it is unclear whether they occur in syntopy. Both can be found in forests and seem to have a preference for the same altered habitats (Lim 1966, 1970; Muul and Liat 1971; Wilson et al. 2006; Wells et al. 2007). Nevertheless, *S. annandalei* is smaller and apparently more arboreal than *S. muelleri*, which is a ground species (although there are some records on trees) often associated with wetter habitats, such as riparian areas (Harrison and Lim 1950; Harrison 1955; Lim 1966, 1970; Muul and Liat 1971). Despite the lack of fine-scale data on the possible syntopy of these 2 species, the differences in ecology, together with morphological divergence in the skull (Figure 4.6) and external morphology (Table 4.3) suggest different niches for *S. annandalei* and *S. muelleri*. However, it is not clear under what ecological conditions *S. annandalei* might have originated, and why it has such a restricted distribution given that there are no apparent ecological barriers to limit its expansion across Sumatra and Peninsular Malaysia.

Morphological divergence in Sundamys

Our morphometric analyses show that lowland and highland taxa occupy different parts of the morphospace for their skull and dentary (Figure 4.6). The lowland *S. muelleri* and *S. annandalei* have larger braincases, larger bullae, and a shorter and broader dentary compared to their highland counterparts (*S. maxi* and *S. infraluteus*). This divergent morphology of the dentary could reflect dietary adaptations, as has previously been found in other Murinae (Satoh 1997; Michaux et al. 2007). Muscle insertions and dentary morphology suggest a higher gnawing capacity (able to process harder food) in both lowland species: *S. annandalei* and *S. muelleri*. On the other side, highland species

with a short in-lever (short condylo-angular distance) and both long molar row and slender jaw (resulting in long out-lever) are likely to be able to close jaws faster, but have less powerful gnawing and chewing capacities. However, the greater molar size provides more chewing surface to process food material in the highland species and this might compensate for this mechanical disadvantage.

Wider and elongated angular process and shorter molar rows, as in the lowland *Sundamys*, have been shown to have a role in gnawing capacities and reduced chewing capacities in some omnivorous rodents (Satoh 1997, Renaud et al. 2007, Samuels 2009) and some herbivorous lineages (*Hapalomys*, *Chiropodomys*, *Pogonomys*, and *Chiruromys*). In some other cases, a larger braincase might be associated with arboreality or to more developed sensory and perceptual capacities, and may indicate the presence of more complex foraging behaviors and more carnivorous or omnivorous diets as compared to folivorous diets (Harvey et al. 1980; Mace et al. 1981). Also, an inflated bullae could be an adaptation to predatory avoidance in lowland *Sundamys*, as has been shown in other rodent communities (Kotler and Brown 1988). More ecological, morphological, and genomic data are needed to better understand the potential roles of divergence and convergence in the evolutionary history of lowland and highland *Sundamys*.

Author contribution: PHF, MCS, and YF obtained the samples. PHF, MCS, and JAL developed the study. PHF, MKT, and MCS did the genetic labwork. PHF did the morphological analysis, and PHF, MCS, and JAL did the phylogenetic analysis. PHF, MCS, and JAL wrote the manuscript. All authors reviewed the manuscript and approved its final version.

Acknowledgements

We thank M. Lakim, F. Tuh Y. Yuh (Sabah Parks, Sabah, Malaysia), M. Hawkins, L. R. Hawkins, F. P. Peter, M. L. Rivera, F. A. C. Marino, and the Malayan porters and field assistants that participated in fieldwork. We thank two anonymous referees for discussions and/or corrections concerning this manuscript. We are grateful to the following people and institutions for granting access to specimens: P. Jenkins, S. Oxford, K. Dixey, and R. P. Miguez (NHM); D. Lunde, N. Edmison, and K. Helgen (NMNH, Smithsonian Institution); E. Westwig, N. Duncan, and R. Voss (AMNH); H.

Baagøe and M. Andersen (ZMUC); G. Véron, V. Nicolas, and C. Denis (MNHN); S. van Der Mije (NBC), J. Woods (DMNH); J. Chupasko (MCZ); Y. Chaval, S. Morand and J. Claude (CBGP); and the Doñana Biological Station (EBD). S.Y.W. Ho provided valuable insight regarding molecular dating. We are grateful to F. Catzeflis for his comments and access to the Montpellier mammal skeleton and tissue collection. We thank A. Mortelliti and R. Castiglia for granting access to tissue of *Paruromys* and *Bunomys*. We thank the team of the CERoPath Project (www.ceropath.org) (and specially the drivers and the students) for sample collection, in particular H. Vibol, K. Aun, K. Blasdell, and P. Buchy in Cambodia, K. Chaisiri in Thailand and Kone in Lao PDR. We are indebted to Serge Morand, Marie Pagès, Julien Claude, and Johan Michaux as the coordinator of the CERoPath project (French ANR Biodiversity, grant ANR 07 BDIV 012) and the BiodivHealthSEA project (www. biodivhealthsea.org) (French ANR CP&ES, grant ANR 11 CPEL 002). Access to the "Plateforme ADN dégradé" (ISEM, UM) was provided by C. Tougard. Logistical support was also provided by Laboratorio de Ecología Molecular, Estación Biológica de Doñana, CSIC (LEM-EBD). We thank the State Ministry of Research and Technology (RISTEK, permit number: 028/SIP/FRP/SMII/2012) and the Ministry of Forestry, Republic of Indonesia for providing permits to carry out fieldwork in Indonesia. Likewise, we thank the Research Center for Biology, Indonesian Institute of Sciences (RCB-LIPI) and the MZB for providing staff and support to carry out fieldwork in the Moluccas. We also thank Sabah Parks (permits: TS/PTD/5/4 Jld. 45 (33) and TS/PTD/5/4 Jld. 47 (25)) for research permits and various kind of support, the Economic Planning Unit (reference: 100-24/1/299), and export permits from the Sabah Wildlife Department (JHL.600-3/7 Jld.7/19 and JHL.600-3/7 Jld.8/) and Sabah Biodiversity Council (Ref: TK/PP:8/8Jld.2). This research received support from the SYNTHESYS Project (http://www.synthesys.info/) which is financed by the European Community Research Infrastructure Action under the FP7 Integrating Activities Program (SYNTHESYS ACCESS GB-TAF-2735, GB-TAF-5026, and GB-TAF-5737 granted to PHF to the NHM, London; NL-TAF-5588 granted to MCS to the NBC, Leiden). The Spanish Ministry of Science and Innovation grants CGL2010-21524 and CGL2014-58793-P also supported this work. MCS is supported by the Spanish Ministry of Science and Innovation Predoctoral Fellowship BES- 2011-049186. PHF was funded by a Marie-Curie fellowship (PIOF-GA-2012-330582-CANARIP-RAT). This publication is

contribution No 2017-XX of the Institut des Sciences de l'Evolution de Montpellier (UMR 5554 – CNRS-IRD).

Literature Cited

ACHMADI, A. A. S., J. A. ESSELSTYN, K. C. ROWE, I. MARYANTO, AND M. T. M. ABDULLAH. 2013. Phylogeny, diversity, and biogeography of Southeast Asian spiny rats (*Maxomys*). Journal of Mammalogy 94:1412–1423.

ADAMS, D. C. AND E. OTÁROLA-CASTILLO. 2013. geomorph: An R package for the collection and analysis of geometric morphometric shape data. Methods in Ecology and Evolution 4:393–399.

APLIN, K. P. ET AL. 2011. Multiple Geographic Origins of Commensalism and Complex Dispersal History of Black Rats. PLoS One 6:e26357.

BENTON, M. J. AND P. C. J. DONOGHUE. 2007. Paleontological evidence to date the tree of life. Molecular Biology and Evolution 24:26–53.

BONHOTE, J. L. 1903. Report on the mammals. Fasciculi Malayenses 1:30–31.

BOOKSTEIN, F. L. 1991. Morphometric tools for landmark data: Geometry and Biology. Cambridge University Press.

BOROWIEC, M. L. 2016. AMAS: a fast tool for alignment manipulation and computing of summary statistics. PeerJ 4:e1660.

BOUCKAERT, R. ET AL. 2014. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. PLoS Computational Biology 10:e1003537.

CHAN, K. L. 1977. Enzyme polymorphism in Malayan rats of the Subgenus *Rattus*. Biochemical Systematics and Ecology 5:161–168.

CHAN, K. L., S. S. DHALIWAL, AND H. S. YONG. 1979. Protein variation and systematics of three subgenera of Malayan rats (Rodentia: Muridae, genus *Rattus* Fischer). Comparative Biochemistry & Physiology 64:329–337.

CHEN, W., Z. SUN, Y. LIU, B. YUE, AND S. LIU. 2012. The complete mitochondrial genome of the large white-bellied rat, *Niviventer excelsior* (Rodentia: Muridae). Mitochondrial DNA 23:363–365.

CLAUDE, J. 2013. Log-Shape Ratios, Procrustes Superimposition, Elliptic Fourier Analysis: Three Worked Examples in R. Hystrix 24:94–102.

CRANBROOK, E. OF, A. H. AHMAD, AND I. MARYANTO. 2014. The mountain giant rat of Borneo *Sundamys infraluteus* (Thomas) and its relations. Journal of Tropical Biology and Conservation 11:49–62.

DRYDEN, I. L. AND K. V. MARDIA. 1998. Statistical Shape Analysis. Wiley, Chichester.

DEMOS, T. C. ET AL. 2016. Local endemism and within-island diversification of shrews illustrate the importance of speciation in building Sundaland mammal diversity. Molecular Ecology 25:5158–5173.

DEN TEX, R.-J., R. THORINGTON, J. E. MALDONADO, AND J. A. LEONARD. 2010. Speciation dynamics in the SE Asian tropics: Putting a time perspective on the phylogeny and biogeography of Sundaland tree squirrels, *Sundasciurus*. Molecular Phylogenetics and Evolution 55:711–20.

ESSELSTYN, J. A., MAHARADATUNKAMSI, A. S. ACHMADI, C. D. SILER, AND B. J. EVANS. 2013. Carving out turf in a biodiversity hotspot: multiple, previously unrecognized shrew species co-occur on Java Island, Indonesia. Molecular Ecology 22:4972–4987.

FABRE, P. ET AL. 2013. A new genus of rodent from Wallacea (Rodentia: Muridae: Murinae: Rattini), and its implication for biogeography and Indo-Pacific Rattini systematics. Zoological Journal of the Linnean Society 169:408–447.

FABRE, P.-H. ET AL. 2016. Mitogenomic phylogeny, diversification, and biogeography of South American spiny rats. Molecular Biology and Evolution 34:613:633.

FABRE, P. H., J. T. VILSTRUP, M. RAGHAVAN, C. D. SARKISSIAN, E. WILLERSLEV, E. J. P. DOUZERY, AND L. ORLANDO. 2014. Rodents of the Caribbean: origin and diversification of hutias unraveled by next-generation museomics. Biology Letters 10:20140266.

HARRISON, J. L. AND B. L. LIM. 1950. Notes on some small mammals of Malaya. Bulletin of Raffles Museum 23:300–309.

HARRISON, J. 1955. The natural food of some rats and other mammals. Bulletin of Raffles Museum 25:157–165.

HARVEY, P. H., T. H. CLUTTON-BROCK, AND G. M. MACE. 1980. Brain Size and Ecology in Small Mammals and Primates. Proceedings of the National Academy of Sciences of the United States of America 77:4387–4389.

HAWKINS, M. T. R., K. M. HELGEN, J. E. MALDONADO, L. L. ROCKWOOD, M. T. N. TSUCHIYA, AND J. A. LEONARD. 2016. Phylogeny, biogeography and systematic revision of plain long-nosed squirrels, (genus *Dremomys*, Nannosciurinae). Molecular Phylogenetics and Evolution 94:752–764.

HEANEY, L. R., D. S. BALETE, E. A. RICKART, M. J. VELUZ, AND S. A. JANSA. 2009. Chapter 7. A New Genus and Species of Small "Tree-Mouse" (Rodentia, Muridae) Related to the Philippine Giant Cloud Rats. Bulletin of the American Museum of Natural History 331:205–229.

HUCHON, D. ET AL. 2002. Rodent phylogeny and a timescale for the evolution of glires: Evidence from an extensive taxon sampling using three nuclear genes. Molecular Biology and Evolution 19:1053–1065.

IUCN. 2015. The IUCN Red List of threatened species. Ver. 2015.3 (www.iucnredlist.org). Accessed 16 December 2015.

JANSA, S. A., F. K. BARKER, AND L. R. HEANEY. 2006. The pattern and timing of diversification of Philippine endemic rodents: evidence from mitochondrial and nuclear gene sequences. Systematic Biology 55:73–88.

JANSA, S. A. AND M. WEKSLER. 2004. Phylogeny of muroid rodents: Relationships within and among major lineages as determined by IRBP gene sequences. Molecular Phylogenetics and Evolution 31:256–276.

JENTINK, F. A. 1879. On some hitherto undescribed species of *Mus* in the Leyden Museum. Notes from the Leyden Museum 2:13–19.

JING, J., X. SONG, C. YAN, T. LU, X. ZHANG, AND B. YUE. 2015. Phylogenetic analyses of the harvest mouse, *Micromys minutus* (Rodentia: Muridae) based on the complete mitogenome sequences. Biochemical Systematics and Ecology 62:121–127.

KATOH, K., K. MISAWA, K. KUMA, AND T. MIYATA. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research 30:3059–3066.

KEARSE, M. ET AL. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28:1647–1649.

KIMURA, Y., M. T. R. HAWKINS, M. M. MCDONOUGH, L. L. JACOBS, AND L. J. FLYNN. 2015. Corrected placement of Mus - Rattus fossil calibration forces precision in the molecular tree of rodents. Scientific Reports 5:14444.

KOTLER, B. P. AND J. S. BROWN. 1988. Environmental heterogeneity and coexistence of desert rodents. Annual Review of Ecology and Systematics 19:281–307.

LANFEAR, R., P. B. FRANDSEN, A. M. WRIGHT, T. SENFELD AND B. CALCOTT. 2016. PartitionFinder 2: New Methods for Selecting Partitioned Models of Evolution for Molecular and Morphological Phylogenetic Analyses. Molecular Biology and Evolution 34:772-773.

LARTILLOT, N., T. LEPAGE, AND S. BLANQUART. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25:2286–2288.

LECOMPTE, E., K. APLIN, C. DENYS, F. CATZEFLIS, M. CHADES AND P. CHEVRET. 2008. Phylogeny and biogeography of African Murinae based on mitochondrial and nuclear gene sequences, with a new tribal classification of the subfamily. BMC Evolutionary Biology 8:199.

LEONARD, J. A., R. J. DEN TEX, M. T. R. HAWKINS, V. MUÑOZ-FUENTES, R. THORINGTON, AND J. E. MALDONADO. 2015. Phylogeography of vertebrates on the Sunda Shelf: A multi-species comparison. Journal of Biogeography 42:871–879.

LI, H. ET AL. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.

LIM, B. L. 1966. Land molluscs as food of Malayan rodents and insectivores. Journal of Zoology 148:554–560.

LIM, B.-L. 1970. Distribution, relative abundance, food habits, and parasite patterns of giant rats (*Rattus*) in West Malaysia. Journal of Mammalogy 51:730–740.

MACE, G. M., P. H. HARVEY, AND T. H. CLUTTON-BROCK. 1981. Brain size and ecology in small mammals. Journal of Zoology 193:333–354.

MARICIC, T., M. WHITTEN, AND S. PÄÄBO. 2010. Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. PLoS One 5:e14004.

MARTIN, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17:10–12.

MCCOMISH, B. J. 2012. Exploring biological sequence space: selected problems in sequence analysis and phylogenetics. PhD dissertation, Massey University, New Zealand.

MICHAUX, J., P. CHEVRET, AND S. RENAUD. 2007. Morphological diversity of Old World rats and mice (Rodentia, Muridae) mandible in relation with phylogeny and adaptation. Journal of Zoological Systematics and Evolutionary Research 45:263–279.

MILLER, M. A., W. PFEIFFER AND T. SCHWARTZ. 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. Proceedings of the Gateway Computing Environments Workshop (GCE), 14 Nov 2010, New Orleans, LA:1-8

MORRISON, D. A. 2008. How to summarize estimates of ancestral divergence times. Evolutionary Bioinformatics 2008:75–95.

MUSSER, G. G. 1981. Results of the Archbold expeditions. No. 105. Notes on systematics of Indo-Malayan murid rodents, and descriptions of new genera and species from Ceylon, Sulawesi, and the Philippines. Bulletin of American Museum of Natural History 168:229–334.

MUSSER, G. G. 1986. Sundaic *Rattus*: definitions of *Rattus baluensis* and *Rattus korinchi*. American Museum Novitates 2862:1–24.

MUSSER, G. G. AND M. D. CARLETON. 2005. Superfamily Muroidea. Pp. 894–1531 in Mammal Species of the World: a taxonomic and geographic reference (D. E. Wilson & D. M. Reeder, eds.). 3rd edition. The Johns Hopkins University Press, Baltimore.

MUSSER, G. G. AND C. NEWCOMB. 1983. Malaysian murids and the giant rat from Sumatra. Bulletin of the American Museum of Natural History 174:327–598.

MUUL, I. AND L. B. LIAT. 1971. New locality records for some mammals of West Malaysia. Journal of Mammalogy 52:430–437.

NILSSON, M. A., A. GULLBERG, A. E. SPOTORNO, U. ARNASON, AND A. JANKE. 2003. Radiation of extant marsupials after the K/T boundary: evidence from complete mitochondrial genomes. Journal of Molecular Evolution 57:3–12.

PAGÈS, M. ET AL. 2010. Revisiting the taxonomy of the Rattini tribe: a phylogeny-based delimitation of species boundaries. BMC Evolutionary Biology 10:184.

PAGÈS, M. ET AL. 2013. Cytonuclear discordance among Southeast Asian black rats (*Rattus rattus* complex). Molecular Ecology 22:1019–34.

PAGÈS, M. ET AL. 2015. Molecular phylogeny of South-East Asian arboreal murine rodents. Zoologica Scripta 45:349–364.

PARADIS, E., J. CLAUDE, AND K. STRIMMER. 2004. APE: Analyses of phylogenetics and evolution in R language. Bioinformatics 20:289–290.

PISANO, J. ET AL. 2015. Out of Himalaya: the impact of past Asian environmental changes on the evolutionary and biogeographical history of *Dipodoidea* (Rodentia). Journal of Biogeography 42:856–870.

RAMBAUT, A., M. A. SUCHARD, D. XIE, AND A. J. DRUMMOND. 2014. Tracer v1.6, Available from http://beast.bio.ed.ac.uk/Tracer.

RENAUD, S., P. CHEVRET, AND J. MICHAUX. 2007. Morphological vs. molecular evolution: ecology and phylogeny both shape the mandible of rodents. Zoologica Scripta 36:525–535.

ROBERTS, T. E., H. C. LANIER, E. J. SARGIS, AND L. E. OLSON. 2011. Molecular phylogeny of treeshrews (Mammalia: Scandentia) and the timescale of diversification in Southeast Asia. Molecular Phylogenetics and Evolution 60:358–372.

ROBINS, J. H., P. A. MCLENACHAN, M. J. PHILLIPS, L. CRAIG, H. A. ROSS, AND E. MATISOO-SMITH. 2008. Dating of divergences within the *Rattus* genus phylogeny using whole mitochondrial genomes. Molecular Phylogenetics and Evolution 49:460–6.

ROBINS, J. H., P. A. MCLENACHAN, M. J. PHILLIPS, B. J. MCCOMISH, E. MATISOO-SMITH, AND H. A. ROSS. 2010. Evolutionary relationships and divergence times among the native rats of Australia. BMC Evolutionary Biology 10:375.

ROHLF, F. 2013. tpsDIG2: Digitize landmarks & outlines from image files, scanner, or video. Available online at http://life.bio.sunysb.edu/morph/soft-dataacq.html.

ROHLF, F. AND D. SLICE. 1990. Extensions of the Procrustes method for the optimal superimposition of landmarks. Systematic Biology 39:40–59.

ROWE, K. C., K. P. APLIN, P. R. BAVERSTOCK, AND C. MORITZ. 2011. Recent and rapid speciation with limited morphological disparity in the genus *Rattus*. Systematic Biology 60:188–203.

ROWE, K. C., M. L. RENO, D. M. RICHMOND, R. M. ADKINS, AND S. J. STEPPAN. 2008. Pliocene colonization and adaptive radiations in Australia and New Guinea (Sahul): multilocus systematics of the old endemic rodents (Muroidea: Murinae). Molecular Phylogenetics and Evolution 47:84–101.

RUDD, R. L. 1965. Weight and growth in Malaysian rain forest mammals. Journal of Mammalogy 46:588–94.

RUEDI, M. AND L. FUMAGALLI. 1996. Genetic structure of gymnures (genus *Hylomys*; Erinaceidae) on continental islands of Southeast Asia: historical effects of fragmentation. Journal of Zoological Systematics and Evolutionary Research 34:153–162.

SAMUELS, J. X. 2009. Cranial morphology and dietary habits of rodents. Zoological Journal of the Linnean Society 156:864–888.

SATOH, K. 1997. Comparative functional morphology of mandibular forward movement during mastication of two murid rodents, *Apodemus speciosus* (Murinae) and *Clethrionomys rufocanus* (Arvicolinae). Journal of Morphology 231:131–141.

SCHENK, J. J., K. C. ROWE, AND S. J. STEPPAN. 2013. Ecological opportunity and incumbency in the diversification of repeated continental colonizations by muroid rodents. Systematic Biology 62:837–864.

SHARIFF, S. M. 1990. Ectoparasites of small mammals trapped at the Ulu Gombak Forest, Selangor Darul Ehsan. The Journal of the Wildlife and Parks IX:9–17.

SIKES, R. S., W. L. GANNON, AND AMERICAN SOCIETY OF MAMMALOGISTS. 2011. Guidelines of the American Society of Mammalogists for the use of wild mammals in research. Journal of Mammalogy 92:235–253.

SLICE, D. E. 2007. Geometric Morphometrics. Annual Review of Anthropology 36:261–281.

STEPPAN, S., R. ADKINS, AND J. ANDERSON. 2004. Phylogeny and divergence-date estimates of rapid radiations in muroid rodents based on multiple nuclear genes. Systematic Biology 53:533–53.

STEPPAN, S. J., R. M. ADKINS, P. Q. SPINKS, AND C. HALE. 2005. Multigene phylogeny of the Old World mice, Murinae, reveals distinct geographic lineages and the declining utility of mitochondrial genes compared to nuclear genes. Molecular Phylogenetics and Evolution 37:370–88.

TAMURA, K., G. STECHER, D. PETERSON, A. FILIPSKI, AND S. KUMAR. 2013. MEGA6: Molecular evolutionary genetics analysis version 6.0. Molecular Biology and Evolution 30:2725–2729.

TILAK, M. K., F. JUSTY, M. DEBIAIS-THIBAUD, F. BOTERO-CASTRO, F. DELSUC, AND E. J. P. DOUZERY. 2015. A cost-effective straightforward protocol for shotgun Illumina libraries designed to assemble complete mitogenomes from non-model species. Conservation Genetics Resources 7:37–40.

TSANGARAS, K. ET AL. 2014. Hybridization capture using short PCR products enriches small genomes by capturing flanking sequences (CapFlank). PloS One 9:e109101.

WANG, S., H. CONG, L. KONG, M. MOTOKAWA, AND Y. LI. 2015. Complete mitochondrial genome of the greater bandicoot rat *Bandicota indica* (Rodentia: Muridae). Mitochondrial DNA 1736:1–2.

WELLS, K., E. K. V. KALKO, M. B. LAKIM, AND M. PFEIFFER. 2007. Effects of rain forest logging on species richness and assemblage composition of small mammals in Southeast Asia. Journal of Biogeography 34:1087–1099.

WILSON, D. E., K. M. HELGEN, C. S. YUN, AND B. GIMAN. 2006. Small mammal survey at two sites in planted forest zone, Bintulu, Sarawak. Malayan Nature Journal 59:165–187.

YONG, H. SEN. 1969. Karyotypes of Malayan rats (Rodentia-Muridae, genus *Rattus* Fischer). Chromosoma 27:245–67.

YOSIDA, T. H. 1973. Evolution of karyotypes and differentiation in 13 *Rattus* species. Chromosoma 40:285–297.

Appendix 4.1. Illumina library preparation

Modern DNA samples were sonicated in a Bioruptor UCD-200 (Diagenode) in volumes of 100 µl at 20 ng/µl in 1.5 ml Eppendorf tubes with 3-6 cycles of 30s on/off with frequency set to high to target mean fragment sizes of 300 bp, or using an ultrasonic cleaning unit (Elmasonic). Historic samples were not sonicated. We prepared libraries following Meyer and Kircher (2010) using half reactions, with some slight modifications. Briefly, we ligated adaptors to the A-tailed samples annealing the sequences corresponding to the binding sites of the Illumina sequencing primers Reads 1 and 2 (Appendix 4.1-Table 1). After indexing PCRs, libraries coming from museum samples had an excess of adaptors of a size very close to the libraries and had to be purified from agarose gels using ISOLATE II PCR and Gel Kit (Bioline). Compared to Meyer and Kircher (2010) protocol, (1) we did the blunt end repair using DNA End Repair Mix (Invitrogen), (2) the fill-in reaction was heat inactivated (20 min at 80 ºC), and (3) the indexing PCR was done in 25 µl using 2.5 U of AmpliTaq Gold Polymerase (Applied Biosystems), 1x Gold Buffer, 2.5 mM $MgCl_2$, 0.25mM dNTP, 0.5 µM PE 1.0, 10 nM PE 2.0 (Appendix 4.1-Table 1), 6.25 µl adaptor ligated DNA and 0.5 µM indexing primer. The PCR program started with 95 ºC for 5 min, followed by 18-21 cycles of 95 ºC for 45s, 60º for 30s, and 72 ºC for 1 min, with a final step of 72 ºC for 10 min. In this way we obtained samples flanked with adaptors, which we indexed with p5 and p7 indexing oligos (Appendix 4.1-Table 1) using 10 cycles. For modern samples we used 0.4 - 2 µM of annealed adaptors while for museum samples we used 0.2 µM. All cleaning steps were carried out using SPRI beads (Rohland and Reich 2012). For the other libraries, prior to enrichment, they were quantified for total DNA using Nanodrop ND-1000 Spectrophotometer (Thermo Scientific) and for mitochondrial DNA with iTaq Universal SYBR Green Supermix (Bio-Rad) in a Stratagene Mx3005P qPCR system following manufacturer's instructions. As a mitochondrial standard we used an amplicon of around 113 bp from the 12S mitochondrial ribosomal gene, which we amplified with primers 12sf4/12sr4m (Appendix 4.1-Table 1) we designed from rodent sequences. We pooled the libraries based on the qPCR (mitochondrial DNA) or Nanodrop values (total DNA), according to if they were to be enriched in mitochondrial or nuclear markers, respectively. The pools were enriched independently for mitochondrial DNA or nuclear genes following the strategies described in Maricic et al.

(2010) for mitogenomes, and in Fabre et al. (2014) and Peñalba et al. (2014) for nuclear genes.

We produced our own bait from PCR products of complete mitochondrial genomes and fragments of nuclear genes *ghr*, *rbp3*, and *rag1*. We amplified the complete mitochondrial genomes from modern samples of *Rattus* and *Sundamys* in two overlapping fragments of around 9.3 kb and 7.5 kb. We used primers described in Sasaki et al. (2005) for *Rattus* and primers Sun7KF/Sun7KR and Sun9KF/Sun9KR (Appendix 4.1-Table 1) for *Sundamys*. We designed these primers from an unpublished complete mitochondrial genome of *Sundamys muelleri* (Appendix 4.1-Table 1). The long-range PCRs were done using the Expand Long Range dNTPack (Roche) in 25 µl reactions with 1x Expand Long Range Buffer, 0.5 µM each dNTP, 0.3 µM each primer, 3 % DMSO, 1.75 U Expand Long Range Polymerase, and around 20 ng of template. The PCR had an initial denaturation 2 min at 95 ºC, 34 cycles of 10s at 92 ºC, 15 s at 62 ºC, 9.5 min at 68 ºC, increasing the extension time by 20 s per cycle after the 9th cycle, and a final elongation of 15 min at 68 ºC. We generated overlapping amplicons for exon 10 of *ghr*, exon 1 of *rbp3*, and exon 1 of *rag1* using primers GHR10/GHR8 and GHR2/GHR7 (Fabre et al. 2013) for *ghr*, I1/J2 (Fabre et al. 2013) and Rpb3F/Rbp3R modified from Fabre et al. (2013 and 2014; Appendix 4.1-Table 1) for *rbp3*, and Rag1F/Rag1R (Appendix 4.1-Table 1) and Rag1b/S70 (this study/Steppan et al. 2004) for *rag1* (Appendix 4.1-Table 1). PCRs were done for *Sundamys* and *Rattus* samples using around 100 ng of DNA. The PCR reactions were done in 50 ul using 1x Gold Buffer, 5 U AmpliTaq Gold, 0.2 mM each dNTP, 20 µg BSA, 0.2 µM of each primer. The PCR reaction started with a denaturation step 10 min at 95 ºC, followed by 35 cycles of 15 s at 95 ºC, an annealing step with a touchdown from 63 ºC to 55 ºC with a rate of –0.5 ºC/cycle, and 1 min at 72 ºC for extension, and ended after a final extension of 10 min at 72 ºC. The long range mitochondrial PCR products were sonicated. The 9.3 kb and 7.5 kb mitochondrial fragments, and the nuclear amplicons were mixed in equimolarity to have two different set of probes for enrichment. They were biotinylated as in Maricic et al. (2010) to make our own baiting molecules. We enriched our libraries as in Maricic et al. (2010) with some modifications, using bait from *Sundamys* for the *Sundamys* samples and bait from *Rattus* for all the other rat samples. We optimized the protocol to work in 96-well plates. The probes were purified to only retain those molecules that had ligated biotinylated adaptors using M-270 streptavidin beads (Dynal,

Oslo, Norway). After a cleaning step we heated the bait-bead complex from 20 ºC to 71 ºC with an increase of 0.5 ºC/s to break the bonds and release the bait molecules (Holmberg et al. 2005), which were then transferred to a clean tube. The hybridization reaction was carried as in Maricic et al. (2010) but in 30 µl, and consisted of 100 ng of bait and per capture and blocking oligos as in Meyer and Kircher (2010) except for BO 1 and 2, that had to be split in two sequences to accommodate the index sequence of the indexing oligo p5 (Appendix 4.1-Table 1). We added 1.5 µg of repetitive sequence fraction of rat genomic DNA, Rat Hybloc DNA (Applied Genetics Laboratories, Melbourne, FL, US), to each nuclear hybridization to block nonspecific hybridization in sequence elements. The hybridization was carried out using thermocyclers for a total of 30 hours starting at 65 ºC and decreasing 1 ºC every 5 hours. After hybridization, M-270 streptavidin beads were used to immobilize the bait-library complexes and washed several times as in Maricic et al. (2010). Lastly, a heat shock at 95 ºC was applied during 1 min to break the biotin-streptavidin bonds and the eluted libraries were transferred to a clean tube. The enriched libraries were quantified with qPCR using LightCycler 96 DNA Green qPCR Master Mix (Roche) in a Stratagene Mx3005P qPCR system with primers Sol Bridge P5/Sol Bridge P7 (Maricic et al. 2010) following manufacturer's instructions. The enriched libraries were re-amplified 6-18 cycles, according to the results of the previous qPCR with the goal of increasing and even their concentrations before final pooling. The PCR was carried in 25 µl with 1x Gold Buffer, 2.5 U AmpliTaq Gold, 3 mM $MgCl_2$, 0.4 µM Sol Bridge P5, 0.4 µM Sol Bridge P7, 0.8 µg/µl BSA, and 12 µl of enriched library. The PCR reaction started with a denaturation step 10 min at 95 ºC, followed by 6 - 18 cycles of 20 s at 95 ºC, 30 s at 60 ºC, and 1 min at 72 ºC for extension, and ended after a final extension of 5 min at 72 ºC. Re-amplified libraries were quantified with Quant-it PicoGreen dsDNA Assay Kit (Invitrogen) on a QuantiFluorTM-ST fluorometer (Promega, US) and pooled together. We sequenced them on an Illumina Hiseq 2000 with 100 bp Paired-End at Macrogen.

Appendix 4.1-Table 1: Oligonucleotides used in library preparation and enrichment.

Name	Sequence 5' -> 3'
Probe generation	
Rbp3F	ATCGCTTACATCCTCAAGCA
Rbp3R	CCATGATGAGGTGCTCTGTGT
Rag1_F	TTATACACTTCCCCTATCTCKAGC
Rag1_R	ACGTGAGTGGTCCCTTCAC
Rag1b	CTCCTTGCTGCTGACCCTAG
Sun7kF	ATAGCAACATGATGACTACTAGCAAGCC
Sun7KR	CTCCATTTCTCTTGTCCTTTCGTACTGGG
Sun9KF	TAGGGTTTACGACCTCGATGTTGGATCAGG
Sun9KR	TTTGGAGTTGCACCAAGGTTTTTGGTTCC
Library preparation	
R1 adaptor	ACACTCTTTCCCTACACGACGCTCTTCCGATC*T
R2 adaptor	P-GATCGGAAGAGCACACGTCTGAACTCCAGTCAC
Ind. oligo p5	AATGATACGGCGACCACCGAGATCTACAC-index-ACACTCTTTCCCTACACGACGCTCTT
Ind. oligo p7	CAAGCAGAAGACGGCATACGAGAT-index-GTGACTGGAGTTCAGACGTGTGCTCTTCCG
PE 1.0	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC*T
PE 2.0	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T
12sf4	GATACCCCACTATGCTTAGCC
12sr4m	GGATATAAAGTACCGCCAAGTC
Enrichment	
BO1.1_P5.1_F	AATGATACGGCGACCACCGAGATCTACAC-P
BO1.2_P5.2_F	ACACTCTTTCCCTACACGACGCTCTTCCGATCT-P
BO2.1_P5.1_R	GTGTAGATCTCGGTGGTCGCCGTATCATT-P
BO2.2_P5.2_R	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT-P

P: phosphate group
*: phosphorothioate bond

Literature Cited in Appendix 1

FABRE, P.-H. ET AL. 2013. A new genus of rodent from Wallacea (Rodentia: Muridae: Murinae: Rattini), and its implication for biogeography and Indo-Pacific Rattini systematics. Zoological Journal of the Linnean Society 169:408–447.

FABRE, P. H. ET AL. 2014. Rodents of the Caribbean: origin and diversification of hutias unravelled by next-generation museomics. Biology Letters 10:20140266.

HOLMBERG, A. ET AL. 2005. The biotin-streptavidin interaction can be reversibly broken using water at elevated temperatures. Electrophoresis 26:501–510.

MEYER, M., AND M. KIRCHER. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harbor Protocols 2010:pdb.prot5448.

MARICIC, T., M. WHITTEN, AND S. PÄÄBO. 2010. Multiplexed DNA sequence capture of mitochondrial genomes using pcr products. Plos One 5:e14004.

PEÑALBA, J. V ET AL. 2014. Sequence capture using PCR-generated probes: a cost-effective method of targeted high-throughput sequencing for nonmodel organisms. Molecular Ecology Resources 14:1000-1010.

ROHLAND, N., AND D. REICH. 2012. Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. Genome Research 22:939–946.

SASAKI, T. ET AL. 2005. Mitochondrial phylogenetics and evolution of mysticete whales. Systematic Biology. 54:77–90.

STEPPAN, S. J., B. L. STORZ, AND R. S. HOFFMANN. 2004. Nuclear DNA phylogeny of the squirrels (Mammalia: Rodentia) and the evolution of arboreality from c-myc and RAG1. Molecular Phylogenetics and Evolution 30:703–19.

Appendix 4.2. Specimens used for the palatal view of the skull in geometric morphometric analysis

Sundamys infraluteus. Females: BMNH 71.2840, 71.2842, 71.2844; MCZ 36107; MZB 23609; FMNH 108934, 108936; USNM 292759, 292763-4, 292766-8, 292770, 301076

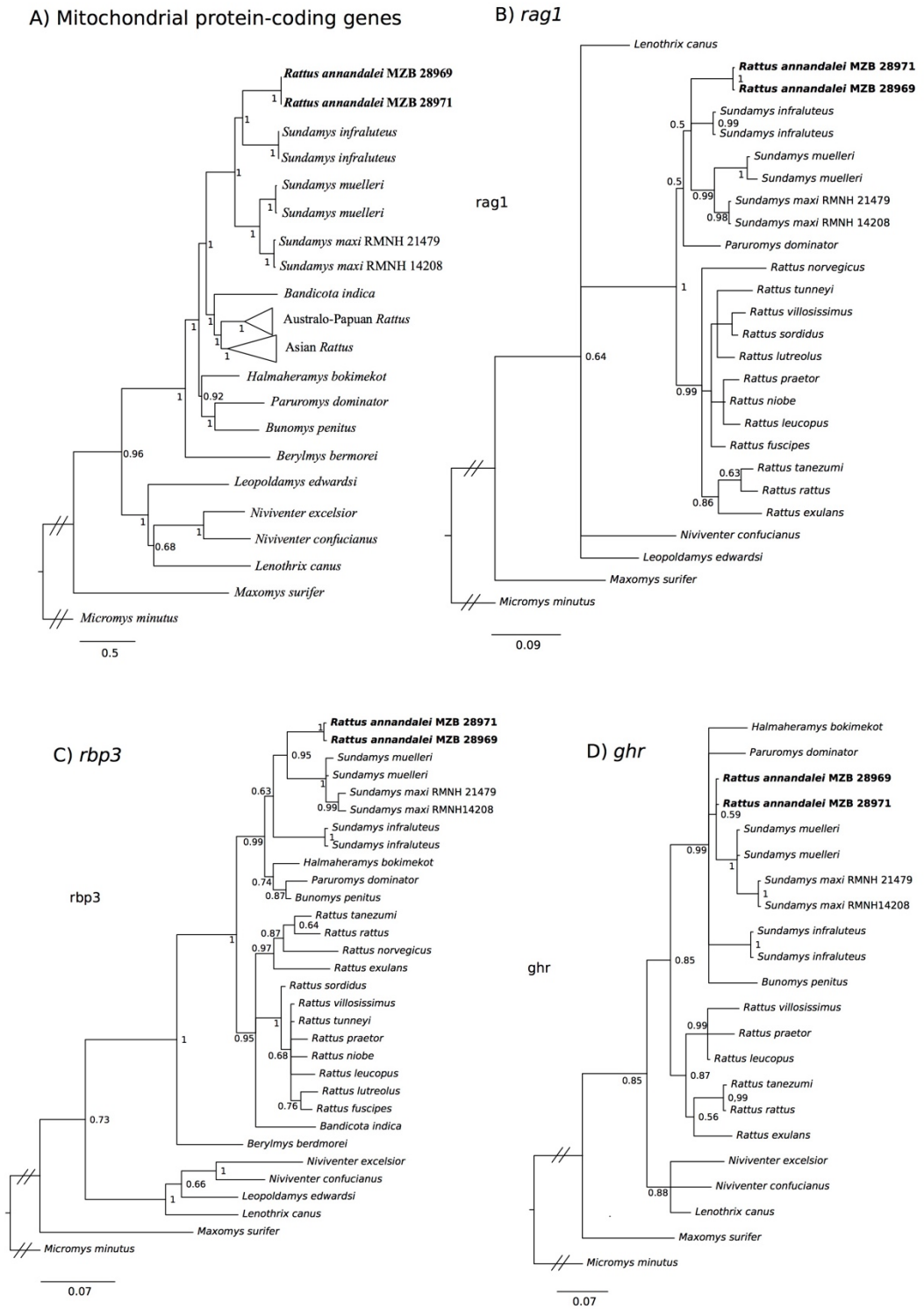
Males: AMNH 106669; BMNH 71.2839, 71.2841; FMNH 108932-3, 108935, 108937; MCZ 36105-6, 36108; USNM 292765, 301075, 301077. *Sundamys maxi*. Females: RMNH 14207, 21479. Males: RMNH 14181, 14205-06, 14210-11, 13566. *Sundamys muelleri*. Females: AMNH 102804-5, 103606; BMNH 9.4.1.437-38, 55.2911, 55.2914, 55.2918, 55.2920; RMNH 21469, 21470, 23279; USNM 104838-39, 113036, 113039, 114286, 114620, 114622, 115585, 115587, 478119, 478121-2, 478128, 478130, 478137, 478139, 478145-46, 478149. Males: AMNH 102547; BMNH 9.4.1.441, 55.2910, 55.2913, 55.2917, 55.939, 55.946; DMNH 6196; FMNH 63154-55, 63157; MNHN CG1977N211, CG1981N258, CG1981N256, CG1990N573; RMNH 18351; USNM 113035, 113052, 114290, 478127, 478129, 478131, 478125. *Sundamys annandalei.* Females: BMNH 553152, 553156; MNHN CG1980N224, CG1981N246, CG1981N248, CG1981N250, CG1981N235-7. Males: BMNH 553153-55; MNHN CG1980N223, CG1980N241, CG1981N233-4, CG1981N238-9, CG1981N242, CG1981N249, CG1981N251, CG1981N254, CG1981N255; MZB 28969. *Rattus baluensis*. Females: FMNH 108908-9, 108911, 108915, 108924; USNM 292696, 292698. Males: BMNH 712771; MZB 5633. *Rattus andamanensis*. Females: USNM 111852, 238174, 279261, 533731, 564485. Males: USNM 111837, 533435-6, 533732-3. *Rattus argentiventer*. Females: MNHN CG1924N281, CG1924N285, CG1969N151. Males: MNHN CG1929N274, CG1957N547, CG1971N808, CG1977N232, CG1977N234, CG1982N101-2. *Rattus exulans*. Females: CBGP L219, L256, K52-53, K62. Males: CBGP L271, K63, K58, K61, R4094. *Rattus losea*. Females: CBGP L268, R4250, R5062, R5190; RMNH 22657. Males: AMNH 275553; CBGP R4791, R4857, R5195; RMNH 22656. *Rattus norvegicus*. Females: CBGP 2446812. Males: CBGP 2446813. *Rattus rattus*. Females: RMNH 9802, 9814. Males: CBGP R121, R129, R132, R148, R149, R152-3; ZMUC 13694. *Rattus tanezumi*. Females: CBGP L8, R4271, R5189, R5422, R5433. Males: CBGP L50, R4114, R4878, R4997; MZB 22715. *Rattus tiomanicus*. Females: USNM 197476-80. Males: MZB 34436; USNM 197481, 197483, 197484.

Appendix 4.3. Specimens used for the dentary lateral side in the morphometric geometric analysis

Sundamys infraluteus. Females: BMNH 712840, 712842, 712844; FMNH 108934, 108936; MCZ 36107; MZB 5084, 23609; RMNH 21253; USNM 292759, 292764, 292766-8, 292770, 301076. Males: BMNH 712839, 712841; FMNH 108932-33, 108935, 108937; MCZ 36105-6, 36108; USNM 292765, 301075, 301077.

Sundamys maxi. Females: RMNH 13566, 13576, 14206-7, 14209, 21479. Males: RMNH 13968, 14181, 14205, 14210-1. *Sundamys muelleri* Females AMNH 103606-7, 103762- 4, 103766, 103768-9; BMNH 552911, 552914, 552918, 552920, 55944-5; MNHN CG1981N259, CG1981N260, CG1990N572; USNM 114286-9, 114291, 114378, 114380, 114382, 478128, 478130, 478137, 478139, 478145. *Sundamys annandalei.* Females: BMNH 32.516, 553157, 553159-60; MNHN CG1980N224, CG1981N235 -7, CG1981N246, CG1981N248, CG1981N250. Males: BMNH 50.952-3, 553158, 611217; MNHN CG1980N241, CG1981N233, CG1981N238-9, CG1981N247, CG1981N249, CG1981N251, CG1981N254-5; MZB 28969.

Appendix 4.4. Per-locus PhyloBayes trees.



Chapter 5 Multilocus nuclear and mitogenome DNA analyses expose complex genetic structure in *Sundamys* rats across Sundaland

Miguel Camacho Sanchez[1], Kristofer M. Helgen[2*,3], Alice Latinne[4*,5], Jesus E. Maldonado[6], Konstans Wells[7] and Jennifer A. Leonard[1]

[1]Conservation and Evolutionary Genetics Group, Estación Biológica de Doñana (EBD-CSIC), Avd. Américo Vespucio 26, 41092 Seville, Spain.

[2 *]current address: School of Biological Sciences, University of Adelaide, Adelaide, South Australia 5005, Australia.

[3]Division of Mammals, National Museum of Natural History, Smithsonian Institution, P.O. Box 37012, Washington DC 20013-7012, USA.

[4*]current address: EcoHealth Alliance, 460 West 34th Street – 17th floor, New York, NY 10001.

[5] Laboratoire de génétique des micro-organismes, Institut de Botanique – B22, Université de Liège, Boulevard du Rectorat, 27 4000, Liège, Belgium.

[6]Smithsonian Conservation Biology Institute, Center for Conservation and Evolutionary Genetics, National Zoological Park, Washington DC 20008, USA.

[7]Environmental Futures Research Institute, School of Environment, Griffith University, Brisbane Qld 4111, Australia.

Abstract

Sundaland is a world biodiversity hotspot. Its complex geological past involves the Sunda Shelf being exposed intermittently during most of the Plio-Pleistoce. This seems to have driven much of the biography in this region, and promote extensive "in-situ" speciation. Vertebrate diversification patterns in this region are starting to be understood. However, most studies in are based on limited taxonomic, geographical or genomic sampling. Here, we sequence 34 nuclear loci and mitogenomes in all species of *Sundamys*, a rat genus endemic to Sunda, including their main populations across the region. We unveil highly structured genetic diversity between main landmasses. Borneo hosts unique non-structured diversity for *S. muelleri*, which is reciprocally monophyletic to its conspecific population from western Sunda, suggesting the presence of an ecological barrier to gene flow in periods of exposed Sunda Shelf. Even, more divergent (8% in *cytb*) are the montane *S. infraluteus* lineages from Borneo and Sumatra, which together with the divergent *S. muelleri* from Palawan (6% *cytb*), merit additional study of their species status.

Introduction

Sundaland is a top world biodiversity hotspot (Myers 2000) delimited by marked zoogeographical barriers. Its relative isolation together with its complex geological history may have contributed to high levels of in-situ speciation (de Bruyn et al. 2014). This is illustrated by a high level of endemism in the vertebrates of Sundaland, around 34% of all species (Myers 2000). The late Pliocene and Pleistocene were characterized by particularly violent climatic oscillations associated with repeated sea level oscillations (Miller et al. 2005) and vegetation changes (Bird et al. 2005; Canon et al. 2009). They created intermittent bridges and vegetation corridors/barriers across the main landmasses which seem to have shaped much of its current vertebrate diversity (Woodruff 2010; Lohman et al. 2011; Leonard et al. 2015; Mason et al. 2016). The effects of this dynamic history can be seen both within and between species.

Intrinsic Sunda diversification dynamics both within and between species are starting to be better understood with the help of genetic tools. These studies have identified general patterns between species (Esselstyn et al. 2009; Leonard et al. 2015; Sheldon et al. 2015; Demos et al. 2016), identified divergent lineages within widespread species (den Tex et al. 2010; Esselstyn et al. 2013), and helped reassign species to their correct genera (Hawkins et al. 2016; Chapter 4). Many phylogeographic or phylogenetic studies of small mammals in the region are hampered due to restricted marker, taxon or geographical sampling (review in Leonard et al. 2015).

The rat genus *Sundamys*, endemic to Sundaland, is an excellent model to study the process of diversification within and between species since the late Pliocene. This genus is distributed across Sundaland and has diversified in this time period (Chapter 4). The process of diversification may proceed differently in species associated with different habitats, which may have reacted differently to environmental changes. This genus includes both narrow endemics and widespread species in different habitats and on multiple landmasses.

Here, we sequenced whole mitogenomes and a panel of 34 nuclear markers from all species and populations on all major landmasses for widely distributed species to provide a phylogeographic framework for this genus. We reveal strong geographically structured cryptic diversity with two or three potential new species.

Methods

Study system

The genus *Sundamys* includes two lowland species: *S. annandalei* and *S. muelleri*. *S. annandalei* was recently moved from *Rattus* to *Sundamys* (Chapter 4), and has a relatively restricted distribution to southern Peninsular Malaysia and eastern Sumatra. *S. muelleri* is very widespread across Sundaland including the Malay Peninsula, Sumatra, Borneo, most small islands and the Palawan transition zone (Esselstyn et al. 2010); but not Java. The other two species, *S. infraluteus*, and *S. maxi*, are tightly associated with montane habitats. *S. infraluteus* is found at high altitudes on Sumatra and Borneo, and *S. maxi* has only been recorded in two locations on Java from 1,000-1,300 m (Musser and Newcomb 1983; Figure 5.1).

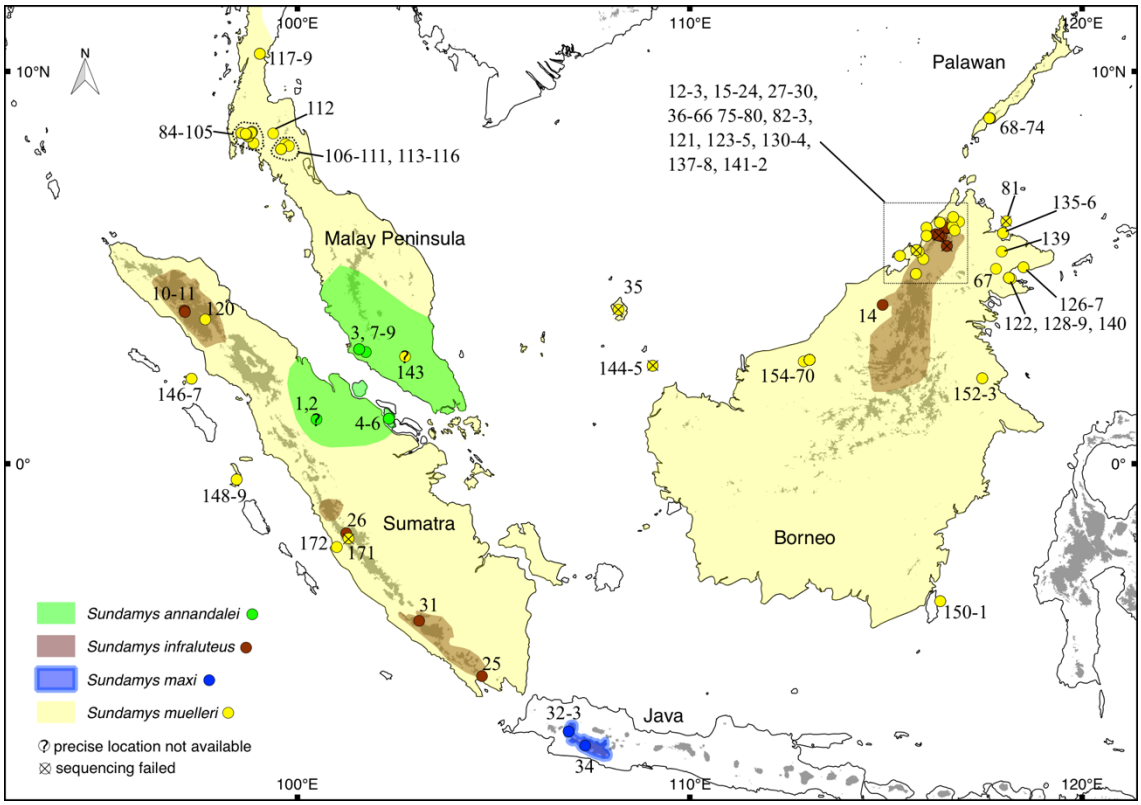


Figure 5-1. IUCN (2015) distribution and sampling of *Sundamys*. Numbers refer to samples in Table 5.1.

Taxon sampling

We sampled 172 individuals of all recognized species in *Sundamys* across their major distribution areas. We included genetic samples from historical and modern animals,

and several GenBank sequences: *S. muelleri*, 7 samples from Palawan, 84 across Borneo, 37 from the Malay Peninsula, 3 from the Natuna Islands, and 7 from Sumatra; *S. infraluteus,* 17 samples from northern Borneo and 5 from Sumatra; *S. annandalei*, 4 samples from Peninsular Malaysia and 5 from the Sumatran side; and *S. maxi*, 3 samples from Java (Figure 5.1; Table 5.1).

Tissue samples were obtained from individuals sampled in the field, and from tissue and dry natural history collections. The modern samples were obtained from animals trapped in Kinabalu National Park, Sabah, Borneo (details in Hawkins 2015), private tissue collections from animals trapped in Thailand (Latinne et al. 2013) and Sabah, Borneo (Wells 2005), tissue collections from the FMNH (*Sundamys muelleri* from Palawan), NMNH (*S. muelleri* from Sarawak, Borneo), and MZB (*S. annandalei* from Sumatra; museum abbreviations in Table 5.1). Individuals from other locations were sampled from historic specimens across several natural history collections (Table 5.1).

Gene sampling

We targeted complete mitochondrial genomes to resolve, which renders excellent support for nodes at the genus level (Robins et al. 2008; Chapter 4). Additionally, four nuclear exons widely used in rodent systematics (recombination activating gene 1, exon 1, *rag1*; retinol-binding protein 3, exon 1, *rbp3*; growth hormone receptor, exon 10, *ghr*; and myc proto-oncogene, exon 2, *c-myc*; Lecompte et al. 2008; Pagès et al. 2010; Fabre et al. 2013), and a panel of 30 introns proposed for proposed to studying phylogenies of closely related mammals (Igea et al. 2010; Table 2.2 in Chapter 2), were genotyped to generate species trees (Edwards 2009).

Molecular methods

DNA was extracted using phenol-chloroform with ethanol precipitation or DNeasy Blood and Tissue Kit (Qiagen). DNA from historic material was extracted in an isolated, dedicated laboratory. All batches of extractions included negatives which were carried through following steps to monitor any possible contamination.

Table 5-1. Samples from all species and geographic regions of *Sundamys* included in this study. Code, is museum or field number; Date collected, is when the specimen was originally taken from the wild; Tissue, is the type of sample used for genetic analysis; Elevation, is in meters above sea level; Locality, is as specific as possible from available associated information; Lat/Long is the latitude/longitude where the sample was collected from the wild; Collector, who originally collected the animal from the wild.

	Code[a]	Date collected[b]	Tissue[c]	Elevation[d]	Locality[e]	Lat/Long[f]	Collector
	Sundamys annandalei						
1	MZB 28969	-	modern	-	Sumatra	-	-
2	MZB 28971	-	"	-	"	-	-
3	NHMR 999000002109	1.May.1985	toe	-	Thai-Malay Peninsula: Selangor: Jenderam Ulu	2.85,101.74	-
4	USNM 143449	31.Mar.1906	crusties	27	Sumatra: P. Padang	1.15,102.33	Abbott, W. L.
5	USNM 143451	1.Apr.1906	"	27	"	"	"
6	USNM 143452	2.Apr.1906	skin	27	"	"	"
7	USNM 488867	29.Mar.1971	"	15	Thai-Malay Peninsula: Selangor: Kuala Langat FR	2.92,101.57	Muul, I. & L.B. Liat
8	USNM 488869	30.Mar.1971	"	15	"	"	"
9	USNM 488872	5.Apr.1971	"	15	"	"	"
	Sundamys infraluteus						
10	ANSP 20355[g]	13.Apr.1939	skin	2073	Sumatra: Aceh Prov.: G. Leuser: Blang-beke camp	3.91,97.12	F.A. Ulmer, Jr
11	ANSP 20356	26.Apr.1939	crusty+skin	2408	Sumatra: Aceh Prov.: G. Leuser: Bivouac 5	3.87,97.13	"
12	B09165*	-	modern	1830	Borneo: Sabah: Crocker Range Park:G. Alab	5.82,116.34	K. Wells
13	BM 1971.2846	22.Jul.1953	-	1510	Borneo: Sabah: G. Trusmadi: Pampang camp	5.55,116.55	J.L. Harrison
14	BM 1978.1549	6.Sep.1977	-	1850	Borneo: Sarawak: Melinau: G. Mulu: camp 4	4.05,114.91	Medway
15	BOR251 (EBD)	24.Feb.2013	liver	1538	Borneo: Sabah: G. Kinabalu: Kinabalu Park HQ	6.01,116.55	M.T.R. Hawkins
16	BOR253 (EBD)	"	"	1538	"	"	"
17	BOR272 (EBD)	27.Feb.2013	"	1593	"	6.00,116.54	"
18	BOR282 (EBD)	28.Feb.2013	"	1538	"	6.01,116.55	
19	BOR510 (EBD)	2.Mar.2013	"	2029	Borneo: Sabah: G. Tambuyukon: Jeneral Camp	6.21,116.7	C.S., Miguel
20	D66*	-	modern	1194	Borneo: Sabah: Mahua, Crocker Range, Borneo	5.80,116.41	K. Wells
21	NH 1984	16.Aug.1971	-	772	Borneo: Sabah: Penampang: Kg. Togudon	5.85,116.27	Hendry Tsen
22	NH 1985	"	-	772	"	"	"
23	NH 1986	11.Dec.1971	-	772	"	"	"
24	NH 1987	16.Aug.1971	-	772	"	"	Hendry Tsen
25	RMNH 21253	1935	-	700	Sumatra: Lampung Prov.: G. Tanggamoes	-5.41,104.7	Max Bartels Jr.
26	RMNH 21254	7.Aug.1918	-	1600	Sumatra: West Sumatra: G. Kerinci: Sg. Kumbang	-1.77,101.24	E. Jacobson
27	S09136*	-	-	1830	Borneo: Sabah: G. Alab	5.82,116.34	K. Wells
28	USNM 301077	21.Jul.1953	skin	1510	Borneo: Sabah: G. Trusmadi, Pampang Camp	5.55,116.55	Elbel, R. E.

	Code[a]	Date collected[b]	Tissue[c]	Elevation[d]	Locality[e]	Lat/Long[f]	Collector
29	ZRC 3045	-	toepad	914	Borneo: Sabah: G. Kinabalu: Kiau	6.02,116.49	-
30	ZRC 4169	-	"	1006	Borneo: Sabah: G. Kinabalu: Kenokok	6.02,116.54	-
31	AMNH 106669	31.Jul.1936	tissue	1800	Sumatra: G. Dempo	-4.00,103.1	J. J. Menden
	Sundamys maxi						
32	RMNH 13968	1934-5	-	1000	Java: southwestern slopes of G. Pangrango-Gede	-6.82,106.91	Max Bartels Jr.
33	RMNH 14208	1934-5	-	1000	"	"	"
34	RMNH 21479	1932	-	1350	Java: Bandung: Ciboeni	-7.18,107.33	Sody
	Sundamys muelleri						
35	BM 1947.1459[h]	7.Sep.1928	-	24	Natuna Is: P. Bunguran	3.93,108.18	F.N. Chasen
36	BOR053 (EBD)	7.Jul.2012	muscle	870	Borneo: Sabah: G. Tambuyukon: Kepuakan camp	6.21,116.7	M.T.R. Hawkins
37	BOR172 (EBD)	7.Aug.2012	liver	334	Borneo: Sabah: G. Tambuyukon: Monggis Substation	6.2,116.75	"
38	BOR173 (EBD)	"	muscle	334	"	"	"
39	BOR410 (EBD)	24.Mar.2013	liver	513	Borneo: Sabah: Poring Hot Springs	6.05,116.7	"
40	BOR411 (EBD)	"	"	542	"	"	"
41	BOR414 (EBD)	"	"	516	"	"	"
42	BOR424 (EBD)	25.Mar.2013	"	521	"	"	"
43	BOR445 (EBD)	27.Mar.2013	"	525	Borneo: Sabah: Poring Hot Springs: Langanan Trail	"	"
44	BOR447 (EBD)	"	"	543	"	"	"
45	BOR448*	"	ear	526	"	"	"
46	BOR561 (EBD)	14.Mar.2013	liver	347	Borneo: Sabah: G. Tambuyukon: Monggis Substation	6.2,116.75	C.S., Miguel
47	BOR562 (EBD)	"	"	347	"	"	"
48	BOR564 (EBD)	"	"	351	"	"	"
49	BOR566 (EBD)	"	"	353	"	"	"
50	BOR567 (EBD)	"	"	331	"	"	"
51	BOR568*	"	ear	351	"	"	"
52	BOR569*	"	"	351	"	"	"
53	BOR571 (EBD)	15.Mar.2013	liver	351	"	"	"
54	BOR572 (EBD)	"	"	351	"	"	"
55	BOR574 (EBD)	"	"	339	"	"	"
56	BOR575*	"	ear	347	"	"	"
57	BOR576 (EBD)	"	liver	351	"	"	"
58	BOR580*	16.Mar.2013	ear	358	"	"	"
59	BOR581*	"	"	352	"	"	"
60	BOR582*	"	"	347	"	"	"

	Code[a]	Date collected[b]	Tissue[c]	Elevation[d]	Locality[e]	Lat/Long[f]	Collector
61	BOR586*	17.Mar.2013	"	351	"	"	"
62	BOR587*	"	"	340	"	"	"
63	BOR588*	18.Mar.2013	"	347	"	"	"
64	BOR589*	"	"	347	"	"	"
65	BOR592*	19.Mar.2013	"	348	"	"	"
66	BOR594*	"	"	351	"	"	"
67	D68*	-	modern	150	Borneo: Sabah: Danum primary forest, Borneo	4.97,117.80	K. Wells
68	FMNH 195419	30.Jun.2007	"	700	Palawan: Rizal Munic: Mt Mantalingahan	8.81,117.63	H. J. D. Garcia
69	FMNH 195420	5.Jun.2003	"	1300	"	8.81,117.65	D. S. Balete
70	FMNH 195421	16.Jul.2007	"	1100	"	8.81,117.64	"
71	FMNH 195422	17.Jul.2007	"	900	"	"	"
72	FMNH 195424	"	"	900	"	"	"
73	FMNH 195426	22.Jul.2007	"	900	"	"	"
74	FMNH 195428	25.Jul.2007	"	700	"	8.81,117.63	"
75	k62*	-	"	167	Borneo: Sabah: Tumbalang logged forest	6.17,116.86	K. Wells
76	KPM 18595	7.Jun.2005	-	10	Borneo: Sabah: P. Gaya	6.02,116.03	-
77	KPM 18598	"	-	10	"	"	-
78	KPM 18599	"	-	486	Borneo: Sabah: Poring Hot Springs	6.05,116.7	-
79	KPM 18878	20.Jun.2005	-	481	Borneo: Sabah: Tenom District, Purulon	5.22,115.95	-
80	KPM 19216	22.Jun.2005	-	71	Borneo: Sabah: Ulu Membakut TBC, Crocker Range	5.41,115.83	-
81	KPM 19270	"	-	0	Borneo: Sabah: Turtle Is. Park: Selingan Is.	6.18,118.06	-
82	KPM 32654	27.Jun.2005	-	216	Borneo: Sabah: Substation Serinsim	6.29,116.71	-
83	KPM 32655	"	-	216	"	"	-
84	L204*	-	modern	36	Thai-Malay Peninsula: Krabi Prov.	8.17,98.88	Lattine, A.
85	L217*	-	"	37	"	8.34,98.75	"
86	L219*	-	"	43	"	8.39,98.77	"
87	L223*	-	"	30	"	8.34,98.74	"
88	L224*	-	"	69	"	8.45,98.83	"
89	L225*	-	"	69	"	"	"
90	L226*	-	"	72	"	8.46,98.84	"
91	L227*	-	"	72	"	"	"
92	L228*	-	"	72	"	"	"
93	L230*	-	"	86	"	8.44,98.82	"
94	L232*	-	"	86	"	"	"

	Code[a]	Date collected[b]	Tissue[c]	Elevation[d]	Locality[e]	Lat/Long[f]	Collector
95	L233*	-	"	86	"	"	"
96	L234*	-	"	86	"	"	"
97	L236*	-	"	86	"	"	"
98	L237*	-	"	86	"	"	"
99	L239*	-	"	171	Thai-Malay Peninsula: Phang Nga Prov.	8.43,98.57	"
100	L247*	-	"	171	"	"	"
101	L248*	-	"	171	"	"	"
102	L251*	-	"	23	Thai-Malay Peninsula: Krabi Prov.	8.39,98.7	"
103	L253*	-	"	23	"	"	"
104	L256*	-	"	34	"	8.41,98.68	"
105	L257*	-	"	34	"	"	"
106	L265*	-	"	90	Thai-Malay Peninsula: Nakhon Si Thammarat Prov.	8.11,99.73	"
107	L266*	-	"	90	"	"	"
108	L268*	-	"	90	"	"	"
109	L269*	-	"	90	"	"	"
110	L277*	-	"	58	"	8.15,99.69	"
111	L279*	-	"	204	"	8.1,99.78	"
112	L282*	-	"	28	"	8.42,99.38	"
113	L293*	-	"	48	"	8.02,99.58	"
114	L294*	-	"	48	"	"	"
115	L295*	-	"	48	"	"	"
116	L296*	-	"	48	"	"	"
117	L479*	-	"	108	Thai-Malay Peninsula: Chumphon Prov.	10.45,99.04	"
118	L480*	-	"	108	"	"	"
119	L483*	-	"	108	"	"	"
120	MVZ 192235	30.Oct.1999	-	390	Sumatra: Ketambe Research Station	3.68,97.65	Jean P. Boubli
121	NH 15	27.Mar.1976	crusties	440	Borneo: Sabah: Ranau: Kg. Muruk	5.95,116.75	-
122	NH 16	23.Sep.1990	"	103	Borneo: Sabah: Lahad Datu: Baturong Kunak	4.74,118.13	-
123	NH 1988	22.Aug.1971	skin+pads	126	Borneo: Sabah: Ulu Tuaran, Kg. Lebodon	6.15,116.37	Hendry Tsen
124	NH 1989	"	"	126	"	"	"
125	NH 2016	"	-	126	"	"	"
126	NH 2050	28.Nov.1979	crusties	31	Borneo: Sabah: Lahad Datu: Silabukan	5.01,118.5	Raymond Goh
127	NH 2060	29.Nov.1979	skin+pads	31	"	"	"
128	NH 2122	29.Jan.1980	-	16	Borneo: Sabah: Lahad Datu: Madai	4.72,118.18	"

	Code[a]	Date collected[b]	Tissue[c]	Elevation[d]	Locality[e]	Lat/Long[f]	Collector
129	NH 25	23.Sep.1990	crusties	103	Borneo: Sabah: Lahad Datu: Baturong Kunak	4.74,118.13	-
130	NH 2514	19.Jun.1983	"	18	Borneo: Sabah: Kinarut: Kg. Limauan	5.81,116.03	Raymond Goh
131	NH 2517	"	-	18	"	"	-
132	NH 2551	3.Jul.1983	crusties	7	Borneo: Sabah: Membakut: Ulu Mawao	5.45,115.77	Raymond Goh
133	NH 2552	"	"	7	"	"	"
134	NH 2559	"	"	7	"	"	"
135	NH 2816	21.Aug.1989	"	35	Borneo: Sabah: Sandakan	5.88,117.98	"
136	NH 2817	23.Aug.1989	"	35	"	"	"
137	NH 2903	5.Oct.1991	-	7	Borneo: Sabah: Beaufort: Mempakul	5.3,115.35	"
138	NH 2928	9.Oct.1991	-	7	"	"	"
139	NH 3260	19.Feb.1986	crusties	6	Borneo: Sabah: Kinabatangan: Batu Putih	5.41,117.95	"
140	NH 36	25.Sept.1990	"	103	Borneo: Sabah: Lahad Datu: Baturong Kunak	4.74,118.13	-
141	NH 3984	2.Jul.1995	-	1365	Borneo: Sabah:Sipitang: G. Lumaku	4.84,115.76	Jaffit Majuakim
142	NH 4723	18.Apr.1998	skin+pads	1580	"	"	"
143	NHMR 999000001085	21.May.1989	toe	NA	Thai-Malay Peninsula	-	-
144	USNM 104838	18.Jun.1900	skin	0	Natuna Is.: Kepulauan Riau: Natuna Is., Serasan Is.	2.5,109.05	Abbott, W. L.
145	USNM 104839	6.Jun.1900	crusties	0	"	"	"
146	USNM 114379	28.Jan.1902	"	54	Sumatra: Banyak Is. P. Tuangku	2.17,97.3	"
147	USNM 114381	4.Feb.1902	skin	54	Sumatra: P. Banjak: Banyak Is., P. Tuangku	"	"
148	USNM 121764	13.Feb.1903	skin	95	Sumatra: west coast: Batu Is., P. Tanahbala	-0.4,98.45	"
149	USNM 121766	14.Feb.1903	crusties	95	"	"	"
150	USNM 151966	3.Jan.1908	skin+crusties	79	Borneo: Kalimantan: P. Sebuku	-3.5,116.38	"
151	USNM 151971	4.Jan.1908	"	79	"	"	"
152	USNM 199013	11.Oct.1912	crustie	12	Borneo: Kalimantan: Birang (=Berau) river	2.18,117.45	Raven, H. C.
153	USNM 199014	"	"	12	"	"	"
154	USNM 590307	17.Jan.2005	modern	17	Borneo: Sarawak: Bintulu Div.: Ulu Kakas	2.65,113.05	Helgen, K. M.
155	USNM 590310	19.Jan.2005	"	17	"	"	"
156	USNM 590311	"	"	17	"	"	"
157	USNM 590312	"	"	17	"	"	"
158	USNM 590314	1.Jan.2005	"	58	Borneo: Sarawak: Bintulu Div.: Ulu Tatau	2.61,112.9	"
159	USNM 590316	16.Jan.2005	"	17	Borneo: Sarawak: Bintulu Div.: Ulu Kakas	2.65,113.05	Wilson, Don E.
160	USNM 590317	17.Jan.2005	"	17	"	"	"
161	USNM 590318	"	"	17	"	"	"
162	USNM 590723	22.Jan.2007	"	22	"	"	Helgen, K. M.

	Code[a]	Date collected[b]	Tissue[c]	Elevation[d]	Locality[e]	Lat/Long[f]	Collector
163	USNM 590724	24.Jan.2007	"	22	"	"	"
164	USNM 590725	26.Jan.2007	"	22	"	"	"
165	USNM 590726	"	"	22	"	"	"
166	USNM 590727	27.Jan.2007	"	22	"	"	"
167	USNM 590728	30.Jan.2007	"	22	"	"	"
168	USNM 590729	31.Jan.2007	"	22	"	"	"
169	USNM 597808	22.Jan.2007	"	22	"	"	"
170	USNM 597809	29.Jan.2007	"	22	"	"	"
171	ZRC 6028	Jun.1914	toepad	945	Sumatra: West Sumatra: Kerinci Valley, Siulak Deras	-1.9,101.3	Robinson & Kloss
172	ZRC 6403	Jun.1914	"	19	Sumatra: West Sumatra: Pasir Ganting	-2.12,101	"

[a] Code of samples from animals trapped and released in the field (*) or from voucher specimens. Museum abbreviation: AMNH, American Museum of Natural History, New York, USA; ANSP, Academy of Natural Sciences, Philadelphia, USA; BM, Natural History Museum, London, UK; FMNH, Field Museum of Natural History, Chicago, Illinois; KPM, Kinabalu Park Museum, Sabah Parks, Sabah, Malaysia; MBZ, Museum Zoologicum Bogoriense, Cibinong, Indonesia; MVZ, Museum of Vertebrate Zoology, Berkeley, USA; NH, Sabah Museum, Kota Kinabalu, Malaysia; NHMR, Natuurhistorisch Museum Rotterdam, Rotterdam, Netherlands; RMNH Naturalis Biodiversity Center, Leiden, Netherlands; ZRC, Raffles Museum, Singapore. EBD stands for specimens hosted at the Doñana Biological Station, Sevilla (Spain), but which have not yet been catalogued.

[b] Date the voucher specimen or field sample was collected from the wild.

[c] modern refers to samples collected in the field and stored in tissue collections, which implies all samples after year 2000. The rest of samples were taken from dry museum specimens, either from foot pads, dry skins, or small pieces of dry tissue attached to the skull (= crusties).

[d] Elevation obtained from field GPS data, specimen labels, and inferred from Lat/Lon coordinates on a Digital Elevation Model of Sundaland (~9 Km resolution).

[e] Abbreviations: G.: Gunung (=mount); Kg.: Kampung (=village); Sg.: Sungai (=river); P.: Pulau (=island); Province: Prov; Division: Div.

[f] Decimal latitude and longitude coordinates from field GPS data, museum databases and inferred from localities.

[g] holotype of *Sundamys infraluteus atchinus*

[h] holotype of *Sundamys muelleri credulus*

Nuclear markers were sequenced from amplicon libraries in the Roche 454 Jr sequencer or from shotgun libraries in Illumina. Amplicon libraries were from multiplex PCR (Methods in Chapter 2) for sequencing the 30 introns in most modern samples (list in Appendix 5.1). Briefly, we amplified all 30 primer pairs (Table 2.2; Chapter 2) in a multiplex PCR with Multiplex PCR Kit (Qiagen), cleaned the products, and ligated barcodes and Roche adapters in a second PCR with Phusion Master Mix (NEB). Amplicon libraries were sequenced in the Roche 454 Jr. The amplicon sequencing worked for most introns, except in *gabrp-1*, although the relative amplification success depended on the marker and species (Figure 1.3; Chapter 1).

Mitogenomes were sequenced following 3 strategies: 1) amplification in two overlapping fragments with a long-range PCR; 2) enrichment of mitochondrial DNA in shotgun libraries; and 3) direct sequencing of shotgun libraries with no-previous enrichment (details in Methods in Chapters 2 and 4; Appendix 5.1 for list of samples sequenced with each strategy). Before library preparation, long-range PCR products and modern DNA extracts were fragmented by sonication to an average size of around 300 bp. Historical samples were processed in an ancient DNA lab, from DNA extractions to the preparation of the indexing PCRs. We prepared libraries from the historical DNA extracts, the modern sonicated DNA extracts and the sonicated long-range mitochondrial PCR products following Meyer and Kircher (2010) and Illumina Kapa Library Preparation Kit (Kapa Biosystems) for most samples (list in Appendix 5.1). We carried out some modifications described in Chapter 4. We did a dual-indexing protocol (Kircher et al. 2012). A few libraries were sequenced with no previous enrichment (Appendix 5.1). The rest were pooled based on species and concentrations and split for enrichment of complete mitochondrial genomes or 34 nuclear loci in independent hybridization reactions following Maricic et al. (2010) for mitogenomes and strategies in Fabre et al. (2014) and Peñalba et al. (2014) for nuclear DNA (see Chapter 4 for modifications). For the preparation of the probes we used mostly *Sundamys* templates or *Rattus* in case primers did not amplify the template in *Sundamys*. Libraries were sequenced in Illumina MiSeq with 250bp PE in Macrogen, and 100 bp PE Hiseq2500 at John Hopkins University.

Pre-processing of sequencing data

We imported the Roche 454 amplicon sequencing data in Geneious R8.1.5 (Biomatters). Reads below 250 nucleotides (nt) were discarded. Then, sequences were

de-multiplexed with the "Standard and Titanium MIDs" in Geneious, in which we replaced default MIDs names with sample names. Around 1% of the reads were not assigned to any sample and were discarded. De-multiplexed reads were exported in FASTAQ. We trimmed primers and adaptors in 5' and 3' ends with cutadapt 1.8.3 (Martin 2011).

All Illumina data was paired-end. Illumina adaptors were removed using cutadapt in paired-end mode and R1 and R2 reads were merged with PEAR v0.9.6 (Zhang et al. 2014). Then, the assembled and unassembled files were merged. Libraries built from long-range PCR products had primers anchored to the 5' or 3' ends. We removed them with cutadapt.

Assembly of mitogenomes

For each sample, we mapped the clean reads to a mitogenome reference of the same species (GenBank accession: *Sundamys infraluteus* KY464175, *S. muelleri* KY464172, *S. maxi* KY464170, and *S. annandalei* KY464176) using BWA 0.7.12-r1039 MEM algorithm (Li 2013). BWA does not consider circularity of mitogenomes, which leads to coverage flanking coordinates of the reference. For this reason, we realigned some samples with low coverage to circularized references in Geneious. We used SAMtools 1.3 (Li et al. 2009) to remove PCR duplicates and called consensus sequences in Geneious (parameters: minimum 2x and 75% threshold).

Genotyping of nuclear data

We reconstructed reference *Sundamys* sequences for the nuclear loci by an initial mapping in Geneious of several *S. muelleri* samples to a *Rattus norvegicus* (Rnor_5.0) reference. For *gabrp-1*, *sfrs5-1* and *fancg-9*, and the 4 exons we used Rnor_5.0 as reference. In all references, we included flanking primer regions. The clean reads from amplicon and shotgun data were mapped to a multifasta reference with the 34 nuclear loci with BWA MEM algorithm with default parameters except the minimum seed length (*-k*) set to 40 to increase stringency. The resulting SAM alignments were converted to binary (BAM) and reads with mapping quality below 50 (most of them were non-target reads mapping to low-complexity regions as determined empirically) were discarded (*samtools view -hu -q50 -F4*). Alignments were sorted and PCR duplicates were removed with SAMtools (*rmdup*).

We merged BAM alignments of sample replicas from the same or different sequencing strategies with SAMtools. This included merging data from amplicon and shotgun sequences to increase per-locus coverage. We used Picard tools (*AddOrReplaceReadGroups*) (Broad Institute 2016) to add a unique sample tag (SM) to each BAM per sample. Independent replicates were also considered for assessment of the genotyping error. We called variants with GATK 3.6 (McKenna et al. 2010) *HaplotypeCaller* tool and did joint genotyping on the gVCFs with the *GenotypeGVCFs* tool. The resulting VCF was intersected with a set of SNPs called with FreeBayes for the same data. Indels were not considered for downstream analysis to simplify downstream analysis and to prevent from calling variants in error-prone regions associated to homo-polymers, which are frequent in Roche 454 data. We used VCFtools v0.1.12b (Danecek et al. 2011) and *vcflib* (github.com/vcflib/) to filter out low quality variants and genotypes. We generated consensus sequences incorporating the variation in the reference using GATK *FastaAlternateReferenceMaker*. Then, we used BEDtools v2.26.0 (Quinlan and Hall 2010) *genomecov* and *maskasta* functions to mask regions in the consensus sequences for which there was less than 3x coverage in the corresponding BAM alignments.

We phased alleles with PHASE 2.1 (Stephens et al. 2001). First, we filtered out loci with > 10% of nucleotides missing with a script in R 3.2.2 (R Core Team 2015). Then, we prepared the PHASE input file with SeqPHASE1 (Flot 2010). PHASE was run with recommended parameters from SeqPHASE1, and all possible alleles were back transformed to FASTA sequences with SeqPHASE2 (Flot 2010). We only considered phased alleles those for which PHASE assigned a probability above 0.9.

Phylogenetic reconstructions with mitogenomes

We aligned mitogenomes with MAFFT 7.310 (Katoh 2013) online version (http://mafft.cbrc.jp/), with algorithm selection automatically set to FFT-NS-2. The alignments were visually inspected and the genes were translated into amino-acids and inspected for stop codons in Geneious. We removed non-protein coding genes in Geneious and *nd6*, which is in the light strand, was reverse-complemented. The mitogenome matrix had 152 rows (149 ingroup individuals), a length of 11330 nt (69 % of the mitogenome), and 0.4 % missing data, as calculated with AMAS (Borowiec 2016). We included mitogenomes of *Rattus baluensis* KY611361, *R. norvegicus* AJ428514, and *R. exulans* (BOR577, EBD, not yet catalogued) as outgroups. The best

partition scheme was determined with PartitionFinder 2.1.1 (Lanfear et al 2016) using the rcluster and the RAxML algorithms. The output partition scheme was specified as input in RAxML 8 (Stamatakis 2014). It arranged the data into one partition for the 1st and 2nd codon positions of all genes, except positions 2 of *ND3* and *ND4* which had their own partitions, and a 4th partition with 3rd codon position for all genes. The rapid bootstrapping algorithm was run on RAxML, which converged after 199 bootstrap following the extended majority-rule (*-autoMRE*) stopping criterion.

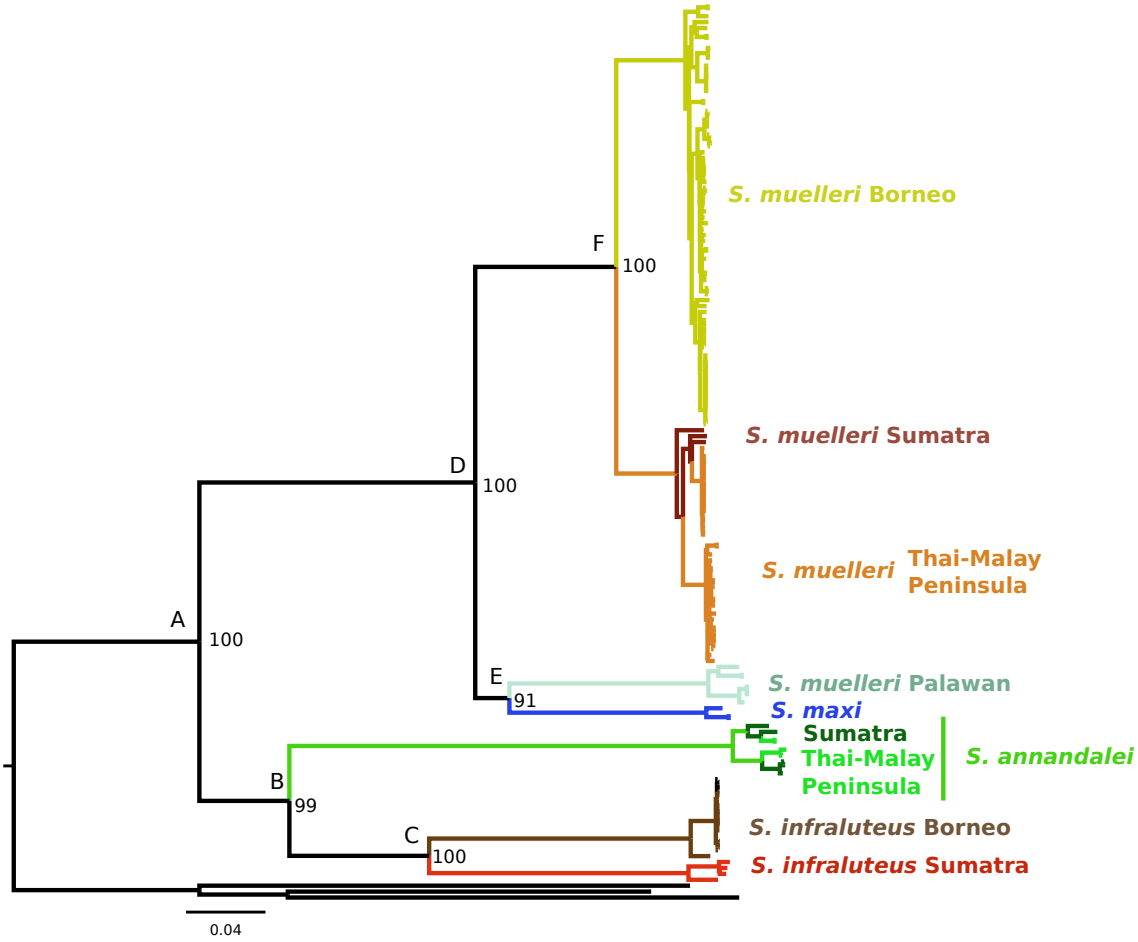


Figure 5-2. Maximum likelihood phylogeny of mitochondrial genome sequences (only protein-coding genes) from all species in the genus *Sundamys* with *Rattus* outgroups (black branches). Key nodes are labeled with bootstrap support and letters for discussion in the text (all samples shown in Appendix 5.2).

Nuclear phylogenetic analysis

We further reconstructed a species tree in a Bayesian framework in *BEAST 2 (Ogilvie et al. 2017). As a preparatory step, evolutionary or "species" groups have to be predefined. To set these groups, we took into consideration the genetic structure in our mitogenome phylogeny (Figure 5.2) and an exploratory multidimensional scaling

(MDS) analysis with the nuclear data (Figure 5.3). Following the genetic structure observed, we defined 6 groups: *S. annandalei*, *S. maxi, S. infraluteus* from Borneo, *S. muelleri* from Palawan, *S. muelleri* from Borneo, and *S. muelleri* from the Thai-Malay Peninsula. We excluded the divergent lineage of *S. infraluteus* from Sumatra due to the high proportion of missing data. We selected up to 8 random alleles per group, choosing preferentially phased alleles. Whenever, phased alleles were not available we phased them manually based on the original BAM alignments or pseudo-phased them by choosing an alternative nucleotide at random for every heterozygous SNP. We excluded *c-myc*, *Rbp3*, *gabrp-1*, and *pipox-5* because we did not have sequences from all lineages, or in the case of *fetub*, because it was probably a case of a duplicated locus (see Chapter 2 and Appendix 5.3). Sequences from introns 14 and 17 of *apeh* were concatenated because they are linked. Thus, the final number of independent loci included in the analysis was 28. We did per locus alignments with MAFFT, and checked them visually in Geneious. The XML BEAST file was prepared in BEAUTi using a STARBEAST2 template. We assumed a strict clock and independent HKY+Γ4+I models for the exonic loci (*ghr* and *rag1*) and the 26 intronic loci, with independent substitution rates. We set all rate exchanges to uniform, a Yule model of speciation, and fixed the root-height of the species tree to an arbitrary value (lognormal distribution: M=2, S=0.0025, in real space). Three chains of 300 million generations were run in BEAST 2.4.6 (Bouckaert et al. 2014). Previous test runs with same parameters but GTR model or model averaging with bModelTEST (Bouckaert and Drummond 2015) for each of these 2 partitions or for each locus had not converge for most parameters of the models.

We evaluated the convergence of the parameters in TRACER v1.6 with a 10% burn-in. All individuals and combined runs showed good posterior convergence. All parameters in the combined log had effective sample sizes (ESS) above 200 except for mutation rates of the exon and intron partitions (100<ESS<200). Invariant sites and gamma shape for the exons *ghr* and *rag1* converged to slightly different values in one of the 3 runs.

The species trees from the 3 runs were combined with LogCombiner 2.4.6 discarding 10% of each tree file, resulting in a total of 162,003 trees. We explored underlying alternative topologies with *treespace* (Jombart et al. 2017) in R. We compared 1000 trees randomly sampled from the 162,003, and determined 4 groups based on the default Kendall Colijn metric, which only considers topology for grouping, but not variation in

branch-lengths. We computed the distance (*treespace::treDist*) between a tree from each group and a 10% of the total trees to determine their group assignation. For each group, we used TreeAnnotator v2.4.4 to determine the maximum clade credibility tree, and FigTree v1.4.2 to produce the tree plots.

Non-parametric genetic structure

We estimated genetic distances using non-parametric methods to describe subjacent genetic structure within complex groups (i.e. *S. muelleri* and *S. infraluteus*) and to highlight genetic distances within and between some of the most divergent lineages we found in our phylogenies.

For the nuclear data, we imported genotypes in the VCF file to R as a *adenet::genind* object (Jombart et al. 2008). We filtered out loci and samples with high proportions of missing data with package *poppr* (Kamvar et al. 2015), only allowing missing data up to 0.3 in individuals and 0.2 in SNPs. We run a MDS analysis with *stats::cmdscale* function in R. We also explored genotype distance, Prevosti's (*poppr::bitwise.dis*) within and between preset groups as defined in the mitochondrial and species tree analysis.

For the mitochondrial DNA, we extracted the *cytb* from the input alignment in RAxML, and computed uncorrected pairwise genetic distances with *ape::dist.dna* (Paradis et al. 2004) in R, for all sequences. Mean and standard deviations were estimated for within-group and between-group comparisons (groups: *S. annandalei*, *S. muelleri* from Thai-Malay Peninsula, *S. muelleri* from Borneo, *S. muelleri* from Palawan, *S. maxi*, *S. infraluteus* from Borneo, *S. infraluteus* from Sumatra, and three *Rattus* species).

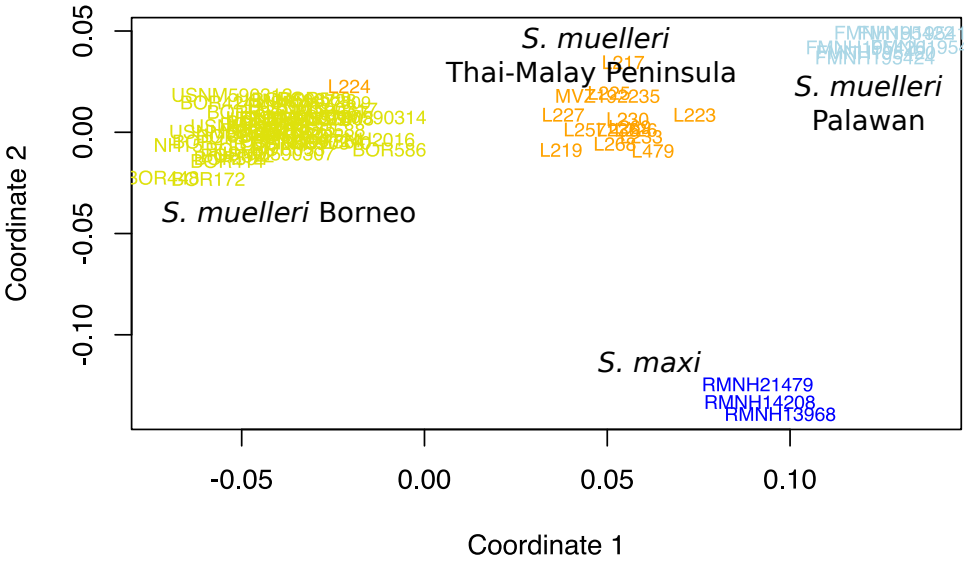


Figure 5-3. Multidimensional scaling (MDS) of the nuclear variation within *Sundamys muelleri* and the closely related *S. maxi*.

Reconstructions of haplotype networks

For nuclear loci, we removed one of each pair of phased alleles if it was considered to be in homozygosis, as determined from the VCF file with an R script. The resulting alleles per locus were converted to nexus format with AMAS (Borowiec 2016). For the mitochondrial data, we extracted the cytochrome *b* (*cytb*) from the mitogenome alignment, removed non-*S. muelleri* samples (*S.m*. from Palawan also removed), and added a GenBank sequence from Kalimantan (DQ191490) and another from Sabah (AM408340). The total matrix contained 113 sequences and 1143 nt. TCS haplotype networks were reconstructed for each *cytb* and nuclear loci in PopArt (Leigh and Bryant 2015).

Results

We reconstructed >90% of the mitogenome for 150 individuals, with an average coverage of 120x. We identified 814 SNPs in 16,330 positions (based on the reference) across the 34 nuclear loci considering all *Sundamys* species. We genotyped >70% of the SNPs in 94 of the 172 individuals (Appendix 5.1). This dataset included all recognized species and samples from all major landmasses for the widespread *S. muelleri*.

Mitochondrial phylogeny and network

The mitochondrial maximum likelihood tree with RAxML had >90% of bootstrap support for principal nodes (nodes A-F; Figure 5.2). Individuals from the Malay Peninsula and Sumatra of both *S. muelleri* and *S. annandalei* did not form reciprocally monophyletic clades. However, *S. muelleri* from western (Sumatra+Malay Peninsula) and eastern (Borneo) Sundaland (node F) formed divergent clades. Even more divergent is the *S. muelleri* lineage from Palawan, which formed a monophyletic clade more related to *S. maxi* than to other *S. muelleri* lineages. The montane *S. infraluteus* from Sumatra and Borneo are deeply divergent sister clades (node C), even deeper than other recognized species: *S. maxi* and *S. muelleri* (nodes D, E). Within Borneo, the southernmost *S. infraluteus*, from Mt. Mulu, was clearly divergent from Sabahan *S. infraluteus*. The lowland species *S. annandalei* is on a long branch more related to the highland *S. infraluteus* (node B) than to the other lowland species, *S. muelleri*.

The *cytb* haplotype network for *S. muelleri* from Borneo and western Sunda individuals (clade F in Figure 5.2), also reflects a profound structure between these groups (Figure 5.4). Unexpectedly, no genetic structure was identified within Borneo or within the western Sunda individuals, with cyt *b*.

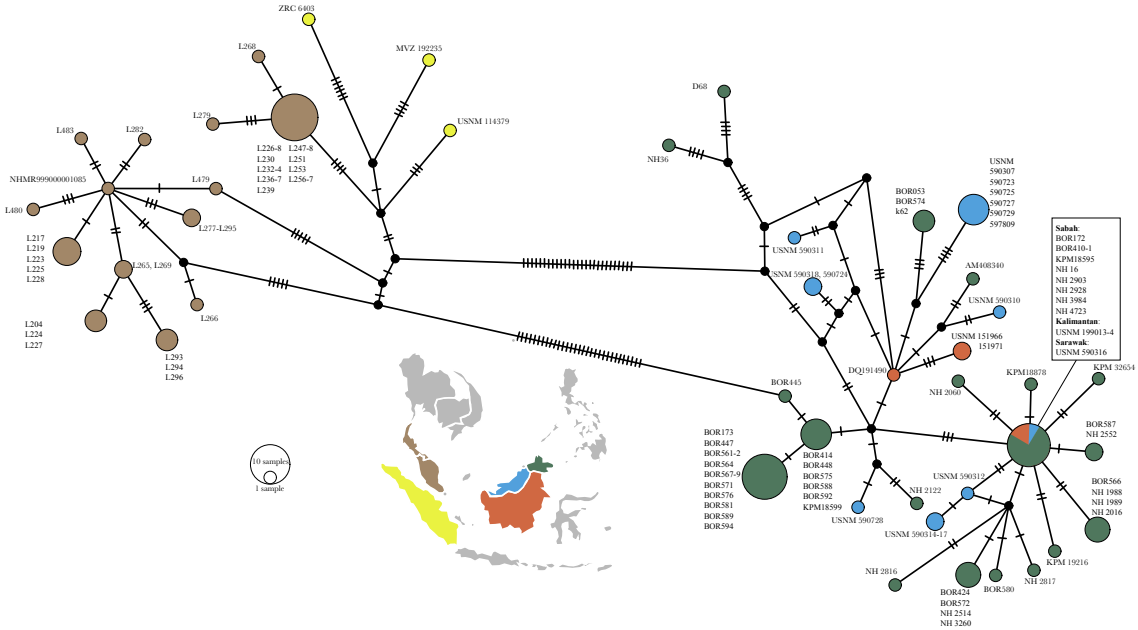


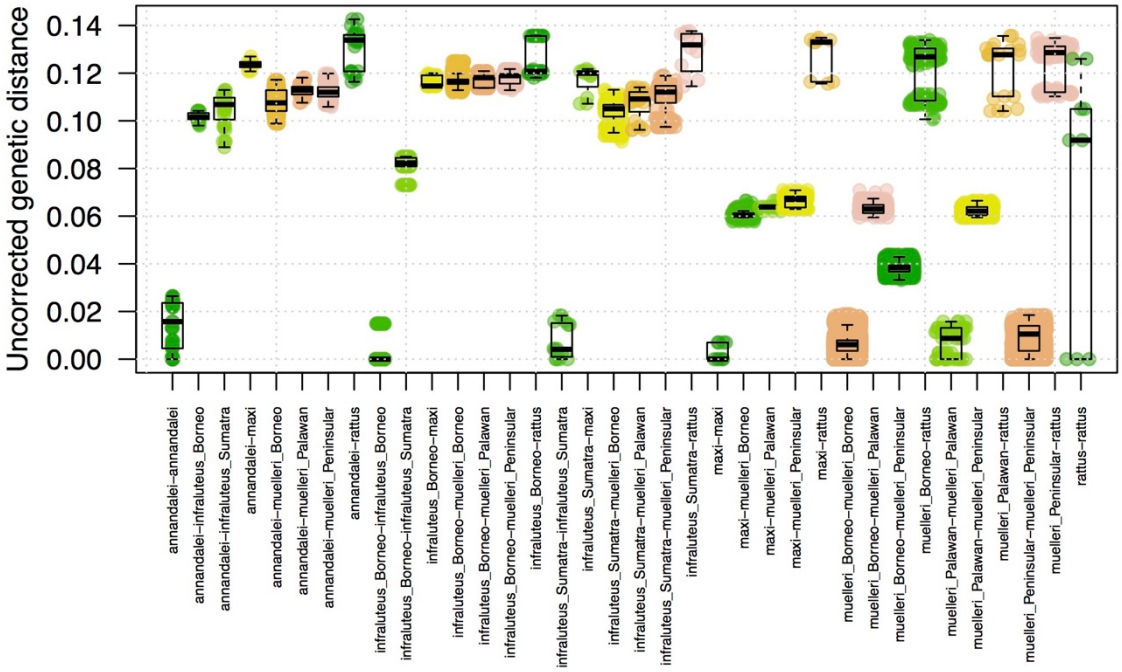
Figure 5-4. Haplotype network of cyt *b* sequences from *Sundamys muelleri* from Borneo, Sumatra and Malay Peninsula.

Genetic differentiation

Most nuclear alleles for *S. infraluteus* were private. However, many alleles were shared between *S. muelleri* across Sumatra, Borneo and Thai-Malay Peninsula. The haplotypes for *S. maxi* and *S. muelleri* from Palawan were mostly peripheral to the diversity in *S. muelleri* from Sumatra, Malay Peninsula and Borneo, and had high proportion of private alleles (Appendix 5.3).

A closer look at the nuclear SNPs with a MDS at the *S. muelleri* complex revealed strong differentiation between *S. muelleri* from Borneo, *S. muelleri* from western Sunda, *S. muelleri* from Palawan, and *S. maxi*. However, as with the mitochondrial data (Figures 5.2 and 5.4), there was no evident substructure within Borneo or western Sunda. Average Prevosti's genotype distances between individuals from different lineages within the *S. muelleri* complex had values between 0.1-0.2, which was higher than variation within groups (up to 0.1), but much lower than distances from Bornean *S. infraluteus* to the others (Figure 5.5 B). Individuals from *S. annandalei* and *S. infraluteus* from Sumatra did not pass missing data thresholds for this analysis. Within-group Prevosti distances were the greatest in *S. muelleri* from Borneo, considering all *Sundamys* lineages.

A)



B)

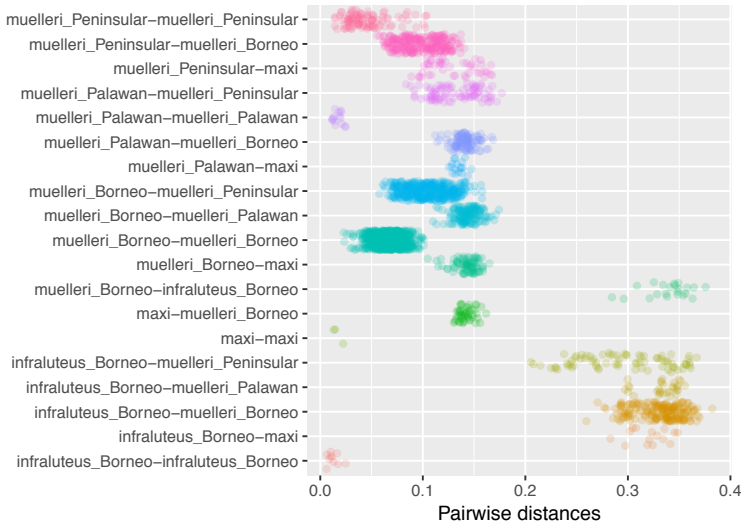


Figure 5-5. A.) *cytb* uncorrected pairwise genetic distances between divergent molecular lineages in *Sundamys*. B.) Pairwise Prevosti's distance between nuclear genotypes for all *Sundamys* lineages (only informative SNPs were kept: minor allele present in at least 2 individuals).

Most within species *cytb* differentiation for across *Sundamys* species was < 2.0% (Figure 5.5; Table 5.2), but that limit was about twice greater (around 4%) for *cytb* distances between Bornean and western Sunda lineages of *S. muelleri*. This differentiation was even greater, around 8%, between *S. infraluteus* from Sumatra and Borneo. This 8% contrasts with the average 6%, between *S. maxi* and other *S. muelleri*

groups, and which is the minimum inters-species distances within *Sundamys*. The divergent population of Palawan also had dissimilarities of around 6% respect to other *S. muelleri* groups (Figure 5.5; Table 5.2).

Table 5-2. Uncorrected mean pairwise genetic distances of *cytb* sequences (%, lower triangular) and their corresponding standard deviations (upper triangular) within and between the molecular mitochondrial groups observed in the mitochondrial in Figure 5.2 (mean/sd, in diagonal).

Molecular group	1	2	3	4	5	6	7	8
1. *annandalei*	1.4/1.0	0.1	0.7	0.2	0.4	0.2	0.3	0.8
2. *infraluteus* Borneo	10.2	0.2/0.5	0.4	0.2	0.2	0.2	0.2	0.7
3. *infraluteus* Sumatra	10.5	8.1	0.7/0.7	0.6	0.5	0.6	0.6	0.9
4. *maxi*	12.3	11.6	11.7	0.3/0.4	0.1	0.1	0.3	0.9
5. *muelleri* Borneo	10.8	11.7	10.3	6.1	0.6/0.4	0.2	0.2	1
6. *muelleri* Palawan	11.3	11.7	10.7	6.4	6.3	0.7/0.6	0.2	1.2
7. *muelleri* Peninsular	11.2	11.8	11	6.6	3.8	6.2	0.9/0.6	0.9
8. *Rattus* spp.	13	12.5	12.9	12.8	12.2	12.2	12.4	7.2/5.5

Multilocus species trees

Phylogenetic inference in *BEAST 2 with 28 nuclear loci converged to posterior species trees with conflicting topologies (Figure 5.6). The mitochondrial topology was only supported by 0.06 % of the trees (Figure 5.6 D). Other trees supported the position of *S. annandalei* in the same clade with the other lowland *Sundamys* (Figure 5.6 A-C). The monophyly of the *S. muelleri-S. maxi* group (muelleri complex hereafter) and the Bornean-Peninsular *muelleri* clade (hereafter core *muelleri*) were supported by all trees (Figure 5.6). *S. maxi* and *S. muelleri* from Palawan had conflicting positions in the tree. Only in 9 % of the trees *S. maxi* was closer to the core *muelleri* group than *S. muelleri* from Palawan (Figure 5.6 B), whereas for the remaining cases, in about half of them (46 %) *S. muelleri* from Palawan and *S. maxi* formed a monophyletic clade (Figure 5.6 A), and for the other half (45 %), the Palawan lineage was closer to core *muelleri* (Figure 5.6 C). In all cases, the branches supporting these alternative topologies within the *muelleri* complex were short, and the 95% of high posterior densities (HPD) of height intervals in these conflicting nodes largely overlapped in all cases..

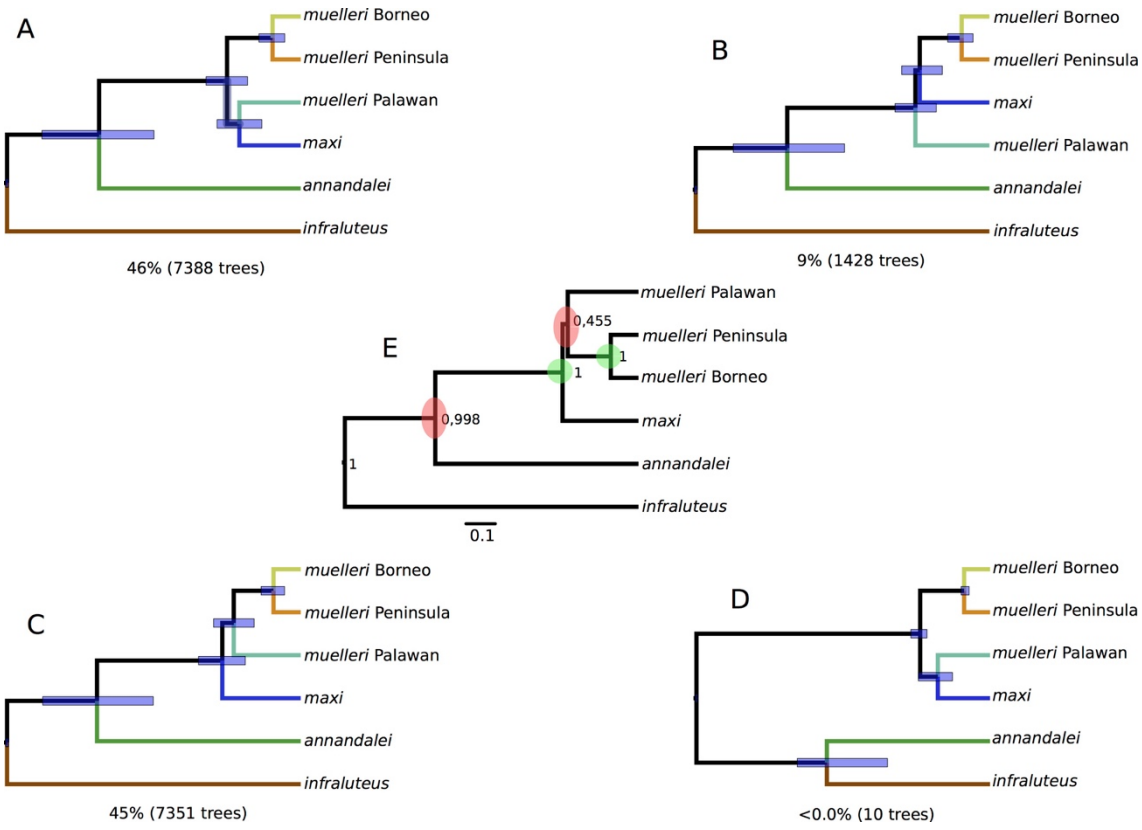


Figure 5-6. Alternative species trees as determined with *treespace* (A-D) and consensus tree (E), inferred with *BEAST 2. Numbers below each A-D tree represent the proportion (%) and total of that topology in the posterior species trees (a subsample of 1/10). All nodes *PP* for A-D is 1.00, but for tree E, where *PP* i indicated in each node and red represents conflicting nodes.

Discussion

Population divergence in lowland species

The lack of genetic structure between populations of *Sundamys annandalei* and *S. muelleri* at both sides of the Malacca Strait (Figure 5.2), separating Sumatra and the Malay-Peninsula, highlight the high permeability of this physical barrier, as already pointed in Leonard et al. (2015). They support the presence of an exposed forested area and gene exchange during low sea level in glacial periods, and the last glacial period in particular, leaving little time for drift into reciprocally monophyletic clades. A quite distinct scenario is observed between the populations of *S. muelleri* from the western Sunda (Sumatra and Thai-Malay Peninsula) and Borneo, which form reciprocally monophyletic clades (Figure 5.2). It rather reflects isolation that predates the last glacial maximum (LGM), as described in other rats (Gorog et al. 2004) and vertebrates (Leonard et al. 2015) with similar distributions. A large block of drier vegetation

prevailed during glaciations conditions in central Sundaland separating western Sunda and Borneo (Bird et al. 2005; Cannon et al. 2009), and it could be responsible for the low permeability of this region to mammals and birds (Sheldon et al. 2015).

The same hypothesis that proposes vicariance during drier conditions in humid forest pockets and connection at rainforest expansion during glacials, has been proposed to explain high rates of endemics and strong genetic structure for Bornean birds (Gawin et al. 2014; Sheldon et al. 2015). However, we found no underlying genetic structure in *Sundamys muelleri* across Borneo, either for nuclear or mitochondrial DNA (Figures 5.2 and 5.3). These results are striking given the sampling spanned over 1,000 km, from Pulau Sebuku, southern Borneo, to Kinabalu National Park, in the north (Figure 5.1), and the high sensitivity of mitochondrial DNA to reflect recent structuration processes (Zink and Barrowclough 2008). This lack of structure could be consequence of the generalist ecology of *S. muelleri* (Payne et al. 2007) combined with its putative large effective population sizes. It further suggests additional barriers to gene flow between *S. muelleri* from different landmasses than mechanisms of isolation-by-distance during the recently exposed Sunda Shelf.

Montane divergence

As we found in high altitude *Rattus*, there is very low diversity within populations of the high altitude *Sundamys infraluteus*, likely reflecting strong drift in small, refugial populations. This should be linked to its reduced suitable habitat (montane forest, mostly associated to streams) and its relative low densities (Nor 2001; Hawkins 2015; field data, unpublished). This habitat specialization is consistent with the greater genetic differentiation between the Sabahan *S. infraluteus* and the individual from Mt. Mulu (Appendix 5.2), the southernmost Bornean distribution for this species (Cranbrook et al. 2014), at around 300 km, connected by a heterogeneous hilly landscape. The lineage from Sumatra was greatly more divergent (around 8% in *cytb*) reflecting the deep history of isolation of these populations. We did not detect any introgression between *S. infraluteus* and the lowland *Sundamys* lineages despite they can be syntopic in an altitudinal range (Musser and Newcomb 1983). This evidence evokes speciation in allopatry (for instance, in forest pockets with forests of colder-like habitats across late Pliocene-early Pleistocene) and later secondary contact, as proposed for birds with similar ecological and genetic patterns along mountains gradients in Borneo (Sheldon et al. 2015; Moyle et al. 2017); and seems contrary to a process of "in-situ" speciation

across an ecological gradient in mountains that supports the origin of *R. baluensis* in Borneo (Chapter 3).

Speciation

We found deep cryptic divergence across *Sundamys*. Mitochondrial DNA supported deeper divergence between *S. infraluteus* from Borneo and Sumatra (deeper node in Figure 5.2, and 8% genetic distance in Figure 5.5 and Table 5.2) than between other recognized species in *Sundamys* (i.e. *S. maxi* with *S. muelleri*). Multilocus nuclear and mitogenome phylogenies highly supported the monophyly of the *muelleri* complex (*S. muelleri* and *S. maxi*), and core *muelleri* (*S. muelleri* from Borneo and western Sunda; Figures 5.2 and 5.6). However, the evolutionary relationships between *S. maxi*, *S. muelleri* from Palawan, and core *muelleri* were not resolved. *Cytb* distances among these 3 lineages are around 6% (Figure 5.5 and Table 5.2), and they had similar branch lengths in the phylogenetic reconstructions with mitogenomes in RAxML (Figure 5.2) and with multilocus nuclear data in *BEAST 2 (Figure 5.6). Genetic distances <2% in *cytb* indicate conspecific individuals (Bradley and Baker 2001, with Kimura 2-parameter model), and divergences >11% are solid for species recognition. The divergences we found between these deep lineages, 6%-8%, fall in between that 2%-11% and merit additional study for species recognition. Morphometrics (Chapter 4), plus discrete and frequency-based traits in the skull (Musser and Newcomb 1983) strongly support the species status of *S. maxi*. Musser and Newcomb (1983) already described smaller size and paler pelage for *S. muelleri* from Palawan, and differences in pelage and some skull proportions for *S. infraluteus* from Borneo and Sumatra.

Our species trees revealed large uncertainties around the evolutionary relationships of *S. maxi*, *S. muelleri* from Palawan, and core *muelleri* (Figure 5.6). These uncertainties, including the cito-nuclear discordances with the mitochondrial data could be ligated to incomplete lineage sorting (Edwards 2008). The short branches supporting the alternative topologies suggest a similar timing of divergence for these lineages, maybe linked to a common event in Sundaland leading *S. muelleri* to invade Java and Palawan. Javanese mammal community is quite distinct from the rest of Sundaland for rodents (Musser and Newcomb 1983) but also for other mammals (Corbet and Hill 1992). Probably the prevailing drier conditions during much of the Pleistocene in central Sunda Shelf and Java made it for forest animals difficult to colonize forest pockets in this island (Sheldon et al. 2015). The origin of *S. maxi* could be correlated to one of the

several waves of colonization in the Pleistocene that characterize Javanese fauna (van der Bergh et al. 2001), and later getting isolated in humid forest, where it was initially described. The divergent lineage of *S. muelleri* from Palawan is part of the characteristic community in this transition zone between Sunda and the Philippines bioregions (Esselstyn et al. 2010). Currently, Borneo and Palawan are separated by a sea floor at -140 m, which is deeper than the sea level drop at the LGM (Voris 2000). The closest event in the last million years to have likely allowed the invasion of Palawan must be traced back at around 430-630 Kya, with a sea level drop of -130 m (review in Esselstyn et al. 2010).

Acknowledgements

We are especially thankful to people that participated in fieldwork in Kinabalu National Park, especially M. T. R. Hawkins. We would like to thank curators and managers that facilitated access to different mammal collections: E. Westwig and N. Duncan, AMNH, New York; N. Gilmore and T. Daeshler, ANSP, Philadelphia; L. Heaney, FMNH, Chicago; F. T. Y. Yu, KPM, Sabah;Y. Fitriana, MBZ, Indonesia ;C. Conroy, MVZ, Berkeley; K. Helgen, N. Edminson, and the Mammal Division at the NMNH, Washington DC; S. van der Mije and P. Kamminga, NBC, Leiden; R. Portela, NHM, London; A. Lo, Sabah Museum, Kota Kinabalu; Kees Moeliker, NMR, Rotterdam;, M. Chua ZRC, Singapore. P.H. Fabre and Y. Fitriana kindly provided DNA extract from MBZ 28969. Logistical support was provided by Laboratorio de Ecología Molecular, Estación Biológica de Doñana, CSIC (LEM-EBD). We also thank Sabah Parks for research permits (TS/PTD/5/4 Jld. 45 (33) and TS/PTD/5/4 Jld. 47 (25)) and various kind of support, the Economic Planning Unit (reference: 100-24/1/299), and export permits from the Sabah Wildlife Department (JHL.600-3/7 Jld.7/19 and JHL.600-3/7 Jld.8/) and Sabah Biodiversity Council (Ref: TK/PP:8/8Jld.2). MCS received support from the SYNTHESYS Project (http://www.synthesys.info/) which is financed by the European Community Research Infrastructure Action under the FP7 Integrating Activities Program: SYNTHESYS ACCESS NL-TAF-5588 to NBC, Leiden, and GB-TAF-5303 to NHM, London. The Spanish Ministry of Science and Innovation grants CGL2010-21524 and CGL2014-58793-P also supported this work. MCS is supported by the Spanish Ministry of Science and Innovation Predoctoral Fellowship BES-2011-049186.

Literature cited

BIRD, M. I. M. I., D. TAYLOR AND C. HUNT. 2005. Palaeoenvironments of insular Southeast Asia during the Last Glacial Period: a savanna corridor in Sundaland? Quaternary Science Reviews 24:2228–2242.

BOROWIEC, M. L. 2016. AMAS: a fast tool for alignment manipulation and computing of summary statistics. PeerJ 4:e1660.

BOUCKAERT, R. AND A. DRUMMOND. 2015. bModelTest: Bayesian phylogenetic site model averaging and model comparison. bioRxiv:20792.

BOUCKAERT, R. ET AL. 2014. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. PLoS Computational Biology 10.

BRADLEY, R. D. AND R. J. BAKER. 2001. A Test of the Genetic Species Concept: Cytochrome-b Sequences and Mammals. Journal of Mammalogy 82:960–973.

BROAD INSTITUTE. 2016. Picard tools. *https://broadinstitute.github.io/picard/*

CANNON, C. H., R. J. MORLEY AND A. B. G. BUSH. 2009. The current refugial rainforests of Sundaland are unrepresentative of their biogeographic past and highly vulnerable to disturbance. Proceedings of the National Academy of Sciences 106:11188–11193.

CORBET, G. B. AND J. E. HILL. 1992. The mammals of the Indomalayan region: a systematic review. P. in. Oxford University Press.

CRANBROOK, E. OF, A. H. AHMAD AND I. MARYANTO. 2014. The mountain giant rat of Borneo *Sundamys infraluteus* (Thomas) and its relations. Journal of Tropical Biology and Conservation 11:49–62.

DANECEK, P. ET AL. 2011. The variant call format and VCFtools. Bioinformatics 27:2156–2158.

DE BRUYN, M. ET AL. 2014. Borneo and Indochina are Major Evolutionary Hotspots for Southeast Asian Biodiversity. Systematic Biology 63:879–901.

DEMOS, T. C. ET AL. 2016. Local endemism and within-island diversification of shrews illustrate the importance of speciation in building Sundaland mammal diversity. Molecular Ecology 25:5158–5173.

DEN TEX, R.-J., R. THORINGTON, J. E. MALDONADO AND J. A LEONARD. 2010. Speciation dynamics in the SE Asian tropics: Putting a time perspective on the phylogeny and biogeography of Sundaland tree squirrels, *Sundasciurus*. Molecular Phylogenetics and Evolution 55:711–20.

ESSELSTYN, J. A, MAHARADATUNKAMSI, A. S. ACHMADI, C. D. SILER AND B. J. EVANS. 2013. Carving out turf in a biodiversity hotspot: multiple, previously unrecognized shrew species co-occur on Java Island, Indonesia. Molecular Ecology 22:4972–4987.

EDWARDS, S. V. 2009. Is a new and general theory of molecular systematics emerging? Evolution 63:1–19.

ESSELSTYN, J. A., C. H. OLIVEROS, R. G. MOYLE, A. T. PETERSON, J. A. MCGUIRE AND R. M. BROWN. 2010. Integrating phylogenetic and taxonomic evidence illuminates complex biogeographic patterns along Huxley's modification of Wallace's Line. Journal of Biogeography 37:2054–2066.

ESSELSTYN, J. A., R. M. TIMM AND R. M. BROWN. 2009. Do geological or climatic processes drive speciation in dynamic archipelagos? The tempo and mode of diversification in Southeast Asian shrews. Evolution 63:2595–2610.

FABRE, P. H. ET AL. 2014. Rodents of the Caribbean: origin and diversification of hutias unraveled by next-generation museomics. Biology Letters 10:20140266.

FABRE, P. ET AL. 2013. A new genus of rodent from Wallacea (Rodentia: Muridae: Murinae: Rattini), and its implication for biogeography and Indo-Pacific Rattini systematics. Zoological Journal of the Linnean Society 169:408–447.

FLOT, J. F. 2010. Seqphase: A web tool for interconverting phase input/output files and fasta sequence alignments. Molecular Ecology Resources 10:162–166.

GAWIN, D. F. ET AL. 2014. Patterns of avian diversification in Borneo: The case of the endemic Mountain Black-eye (*Chlorocharis emiliae*). The Auk 131:86–99.

GOROG, A. J., M. H. SINAGA AND M. D. ENGSTROM. 2004. Vicariance or dispersal? Historical biogeography of three Sunda shelf murine rodents (*Maxomys surifer, Leopoldamys sabanus and Maxomys whiteheadi*). Biological Journal of the Linnean Society 81:91–109.

HAWKINS, M. T. R. 2015. Biogeography and Phylogeography of Mammals of Southeast Asia: A Comparative Analysis Utilizing Macro and Microevolution. Doctoral thesis. George Mason University.

HAWKINS, M. T. R., K. M. HELGEN, J. E. MALDONADO, L. L. ROCKWOOD, M. T. N. TSUCHIYA AND J. A. LEONARD. 2016. Phylogeny, biogeography and systematic revision of plain long-nosed squirrels (genus *Dremomys*, Nannosciurinae). Molecular Phylogenetics and Evolution 94:752–764.

IGEA, J., J. JUSTE AND J. CASTRESANA. 2010. Novel intron markers to study the phylogeny of closely related mammalian species. BMC Evolutionary Biology 10:369.

IUCN. 2015. The IUCN Red List of threatened species. Ver. 2015.3 (www.iucnredlist.org). Accessed 16 December 2015.

JOMBART, T. ET AL. 2008. Package "adegenet". Bioinformatics Application Note 24:1403–1405.

JOMBART, T., M. KENDALL, J. ALMAGRO-GARCIA AND C. COLIJN. 2017. *treespace*: statistical exploration of landscapes of phylogenetic trees. Molecular Ecology Resources.DOI:10.1111/1755-0998.12676

KAMVAR, Z. N., J. F. TABIMA AND N. J. GRÜNWALD. 2014. *Poppr*: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. PeerJ 2:e281.

KATOH, K. AND D. M. STANDLEY. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Molecular Biology and Evolution 30:772–780.

KIRCHER, M., S. SAWYER AND M. MEYER. 2012. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic acids research 40:e3.

LANFEAR, R., P. B. FRANDSEN, A. M. WRIGHT, T. SENFELD AND B. CALCOTT. 2016. PartitionFinder 2: New Methods for Selecting Partitioned Models of Evolution for Molecular and Morphological Phylogenetic Analyses. Molecular Biology and Evolution 34:772-773

LATINNE, A., S. WAENGSOTHORN, P. ROJANADILOK, K. EIAMAMPAI, K. SRIBUAROD AND J. R. MICHAUX. 2013. Diversity and endemism of Murinae rodents in Thai limestone karsts. Systematics and Biodiversity 11:323–344.

LECOMPTE, E., K. APLIN, C. DENYS, F. CATZEFLIS, M. CHADES AND P. CHEVRET. 2008. Phylogeny and biogeography of African Murinae based on mitochondrial and nuclear gene sequences, with a new tribal classification of the subfamily. BMC Evolutionary Biology 8:199.

LEIGH, J. W. AND D. BRYANT. 2015. POPART: Full-feature software for haplotype network construction. Methods in Ecology and Evolution 6:1110–1116.

LEONARD, J. A., R. J. DEN TEX, M. T. R. HAWKINS, V. MUÑOZ-FUENTES, R. THORINGTON AND J. E. MALDONADO. 2015. Phylogeography of vertebrates on the Sunda Shelf: A multi-species comparison. Journal of Biogeography 42:871–879.

LI, H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at arXiv:1303.3997v2 [q-bio.GN]:3.

LI, H. ET AL. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.

LOHMAN, D. J. ET AL. 2011. Biogeography of the Indo-Australian Archipelago. Annual Review of Ecology, Evolution, and Systematics 42:205–226.

MARICIC, T., M. WHITTEN AND S. PÄÄBO. 2010. Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. PLoS One 5:e14004.

MARTIN, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17:10–12.

MASON, V. C. ET AL. 2016. Genomic analysis reveals hidden biodiversity within colugos, the sister group to primates. Science Advances 2:e1600633–e1600633.

MCKENNA, A. ET AL. 2010. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research 20:1297–1303.

MEYER, M. AND M. KIRCHER. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harbor Protocols 2010:pdb.prot5448.

MILLER, K. G. 2005. The Phanerozoic Record of Global Sea-Level Change. Science 310:1293–1298.

MOYLE, R. G., J. D. MANTHEY, P. A. HOSNER, M. RAHMAN, M. LAKIM AND F. H. SHELDON. 2017. A genome-wide assessment of stages of elevational parapatry in Bornean passerine birds reveals no introgression: implications for processes and patterns of speciation. PeerJ 5:e3335.

MUSSER, G. G. AND C. NEWCOMB. 1983. Malaysian murids and the giant rat from Sumatra. Bulletin of the American Museum of Natural History 174:327–598.

MYERS, N., R. A. MITTERMEIER, C. G. MITTERMEIER, G. A. B. DA FONSECA AND J. KENT. 2000. Biodiversity hotspots for conservation priorities. Nature 403:853–8.

NOR, S. M. 2001. Elevational diversity patterns of small mammals on Mount Kinabalu, Sabah, Malaysia. Global Ecology and Biogeography 10:41–62.

OGILVIE, H. A., R. R. BOUCKAERT AND A. J. DRUMMOND. 2017. StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. Molecular Biology and Evolution.

PAGÈS, M. ET AL. 2010. Revisiting the taxonomy of the Rattini tribe: a phylogeny-based delimitation of species boundaries. BMC Evolutionary Biology 10:184.

PARADIS, E., J. CLAUDE AND K. STRIMMER. 2004. APE: Analyses of phylogenetics and evolution in R language. Bioinformatics 20:289–290.

PAYNE, J., C. M. FRANCIS, K. PHILLIPPS AND K. PHILLIPS. 2007. A Field Guide to the Mammals of Borneo. P. in. 3rd edition. book, The Sabah Society, Kota Kinabalu, Sabah.

PEÑALBA, J. V ET AL. 2014. Sequence capture using PCR-generated probes: a cost-effective method of targeted high-throughput sequencing for nonmodel organisms. Molecular Ecology Resources 14:1000–1010.

QUINLAN, A. R. AND I. M. HALL. 2010. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842.

ROBINS, J. H., P. A. MCLENACHAN, M. J. PHILLIPS, L. CRAIG, H. A. ROSS AND E. MATISOO-SMITH. 2008. Dating of divergences within the Rattus genus phylogeny using whole mitochondrial genomes. Molecular Phylogenetics and Evolution 49:460–6.

SHELDON, F. H., H. C. LIM AND R. G. MOYLE. 2015. Return to the Malay Archipelago: the biogeography of Sundaic rainforest birds. Journal of Ornithology. DOI:10.1007/s10336-015-1188-3

STAMATAKIS, A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.

STEPHENS, M., N. J. SMITH AND P. DONNELLY. 2001. A New Statistical Method for Haplotype Reconstruction from Population Data. The American Journal of Human Genetics 68:978–989.

VAN DEN BERGH, G. D., J. DE VOS AND P. Y. SONDAAR. 2001. The Late Quaternary palaeogeography of mammal evolution in the Indonesian Archipelago. Palaeogeography, Palaeoclimatology, Palaeoecology 171:385–408.

VORIS, H. K. 2000. Maps of Pleistocene Sea Levels in Southeast Asia: shorelines, river systems and Time Durations. Journal of Biogeography 27:1153–1167.

WELLS, K. 2005. Impacts of rainforest logging on non-volant small mammal assemblages in Borneo. Doctoral thesis. Universität Ulm.

WOODRUFF, D. S. 2010. Biogeography and conservation in Southeast Asia: how 2.7 million years of repeated environmental fluctuations affect today's patterns and the future of the remaining refugial-phase biodiversity. Biodiversity and Conservation 19:919–941.

ZHANG, J., K. KOBERT, T. FLOURI AND A. STAMATAKIS. 2014. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics 30:614–620.

ZINK, R. M. AND G. F. BARROWCLOUGH. 2008. Mitochondrial DNA under siege in avian phylogeography. Molecular Ecology 17:2107–2121.

Appendix 5.1. Sequencing data.

Sequencing strategy and output for mitogenomes and nuclear loci.

		Mitogenomes			34 nuclear loci	
	Code	Coverage	Seq. strategy*	% reconstructed	Seq. strategy*	% SNPs genotyped
1	MZB 28969	99.5	KY464176*	100.0	ESL	89.4
2	MZB 28971	5	KY464177*	95.6	-	KY467076, -85-6*
3	NHMR 999000002109	39.1	ESL	100.0	"	7.9
4	USNM 143449	139.2	"	100.0	"	0.0
5	USNM 143451	258.5	"	100.0	"	11.1
6	USNM 143452	15.9	"	96.9	"	8.2
7	USNM 488867	62.7	"	100.0	"	28.9
8	USNM 488869	108.8	"	100.0	"	41.8
9	USNM 488872	43.1	"	100.0	"	19.8
10	ANSP 20355	3.4	"	67.2	"	1.8
11	ANSP 20356	1.9	"	47.2	"	0.5
12	B09165*	100	"	100.0	"	47.9
13	BM 1971.2846	0.087	"	0.8	"	0.0
14	BM 1978.1549	51.2	"	100.0	"	74.2
15	BOR251 (EBD)	46	KY464174	100.0	454AS, ESL	95.5
16	BOR253 (EBD)	41	ESL	100.0	"	92.4
17	BOR272 (EBD)	62.6	"	100.0	"	87.2
18	BOR282 (EBD)	11.7	NESL	99.8	454AS	69.4
19	BOR510 (EBD)	142.7	KY464175	100.0	454A, ESL	95.5
20	D66*	129.3	ESL	100.0	ESL	86.1
21	NH 1984	70.5	"	100.0	"	2.7
22	NH 1985	47.3	"	100.0	"	82.1
23	NH 1986	28	"	99.7	"	39.3
24	NH 1987	52	"	100.0	"	9.2
25	RMNH 21253	13.1	"	99.5	"	60.3
26	RMNH 21254	22.2	"	99.4	"	2.0
27	S09136*	109.1	"	100.0	"	83.5
28	USNM 301077	0	"	0.5	"	0.0
29	ZRC 3045	69.2	"	100.0	"	71.7
30	ZRC 4169	0.65	"	13.9	"	0.0
31	AMNH 106669	4.3	"	83.1	"	0.6
32	RMNH 13968	48.5	"	100.0	"	95.9
33	RMNH 14208	51.3	KY464171	100.0	"	95.6
34	RMNH 21479	55.7	KY464170	99.8	"	95.8
35	BM 1947.1459	0.043	ESL	1.0	"	0.0
36	BOR053 (EBD)	411.9	454LR	100.0	454AS	73.5
37	BOR172 (EBD)	43.1	ESL	100.0	454AS, ESL	80.7
38	BOR173 (EBD)	489.3	454LR	100.0	454AS	74.1
39	BOR410 (EBD)	79.7	KY464173	100.0	454AS, ESL	97.8
40	BOR411 (EBD)	431.4	IlLR	100.0	454AS	74.9
41	BOR414 (EBD)	60.6	NESL	100.0	"	72.9
42	BOR424 (EBD)	426	IlLR	100.0	"	71.1
43	BOR445 (EBD)	77.8	ESL	100.0	454AS, ESL	95.2
44	BOR447 (EBD)	300.8	IlLR	100.0	454AS	74.1
45	BOR448*	99.8	KY464172	100.0	454AS, ESL	96.3
46	BOR561 (EBD)	178.3	ESL	100.0	"	98.2
47	BOR562 (EBD)	455	IlLR	100.0	454AS	71.1
48	BOR564 (EBD)	36.4	ESL	100.0	454AS, ESL	82.8
49	BOR566 (EBD)	395.6	IlLR	100.0	-	-
50	BOR567 (EBD)	480.1	"	100.0	454AS	70.6
51	BOR568*	421.8	"	100.0	"	73.2

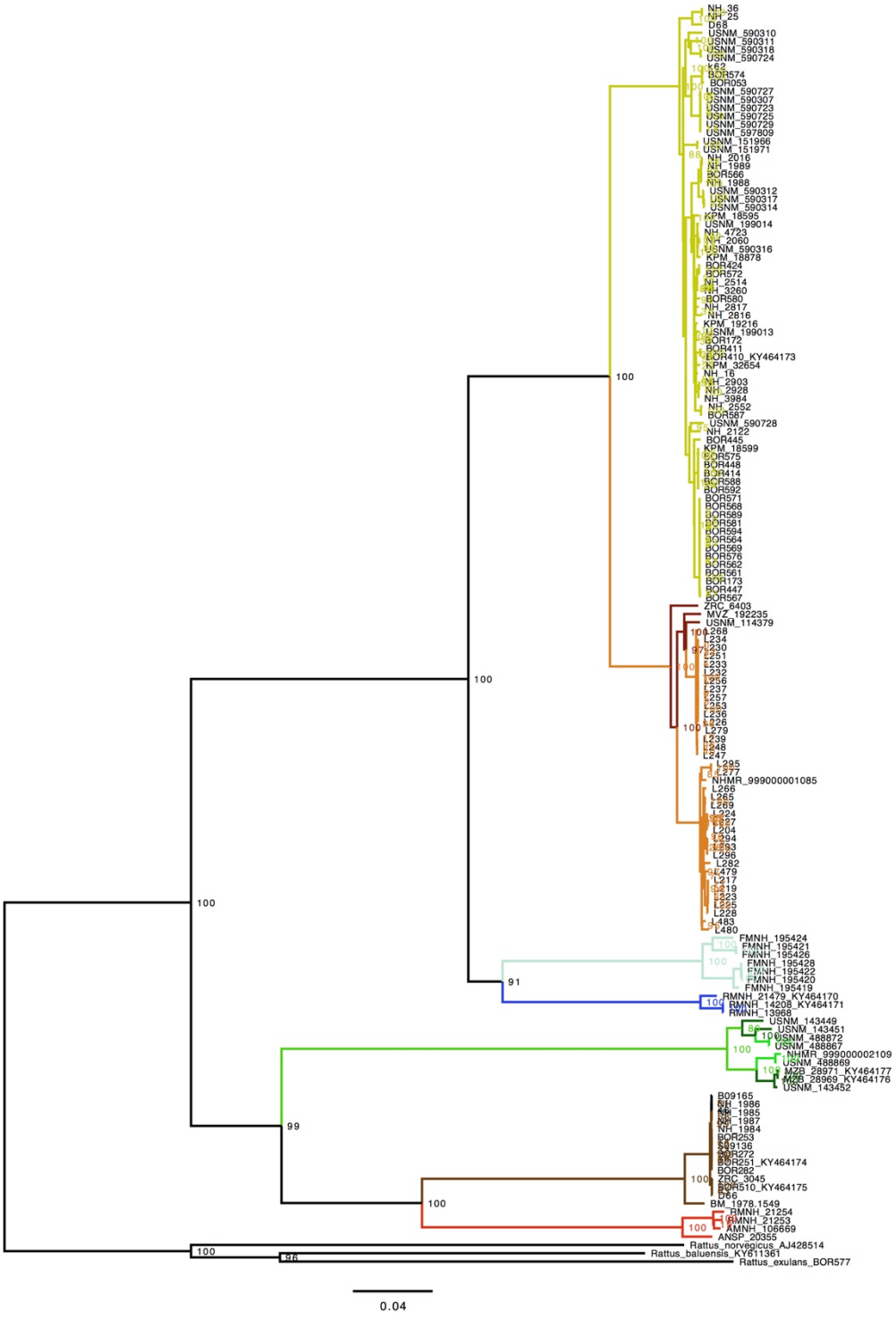
		Mitogenomes			34 nuclear loci	
	Code	Coverage	Seq. strategy*	% reconstructed	Seq. strategy*	% SNPs genotyped
52	BOR569*	37.4	ESL	99.9	454AS, ESL	95.5
53	BOR571 (EBD)	150.4	"	100.0	-	-
54	BOR572 (EBD)	14.2	NESL	100.0	454AS	70.1
55	BOR574 (EBD)	198.9	"	100.0	"	73.1
56	BOR575*	144.5	ESL	100.0	"	74.6
57	BOR576 (EBD)	413.8	IlLR	100.0	"	-
58	BOR580*	92.8	ESL	100.0	454AS, ESL	65.2
59	BOR581*	94.7	"	100.0	"	80.1
60	BOR582*	-	-	-	454AS	72.9
61	BOR586*	-	-	-	"	67.9
62	BOR587*	420	IlLR	100.0	"	69.9
63	BOR588*	67.4	ESL	100.0	454AS, ESL	80.1
64	BOR589*	50.7	"	100.0	"	86.2
65	BOR592*	48.9	"	100.0	"	85.6
66	BOR594*	33.7	"	100.0	"	85.3
67	D68*	157.5	"	100.0	ESL	53.6
68	FMNH 195419	65.5	"	100.0	454AS, ESL	92.4
69	FMNH 195420	236.8	"	100.0	"	77.8
70	FMNH 195421	227.1	"	100.0	"	73.5
71	FMNH 195422	372	IlLR	100.0	454AS	68.9
72	FMNH 195424	236.8	ESL	100.0	454AS, ESL	81.1
73	FMNH 195426	198.9	"	100.0	"	85.5
74	FMNH 195428	228.2	"	100.0	-	-
75	k62	163.7	"	100.0	ESL	41.0
76	KPM 18595	20.4	"	99.7	"	0.0
77	KPM 18598	0.24	"	2.6	"	0.2
78	KPM 18599	55.4	"	99.9	"	0.7
79	KPM 18878	122.6	"	100.0	"	3.7
80	KPM 19216	56.2	"	100.0	"	16.3
81	KPM 19270	1.2	"	30.7	"	5.9
82	KPM 32654	105.9	"	100.0	"	3.4
83	KPM 32655	0.033	"	0.0	"	0.0
84	L204*	98.4	"	100.0	"	84.3
85	L217*	77.6	"	100.0	"	92.4
86	L219*	77.4	"	100.0	"	92.1
87	L223*	84.8	"	100.0	"	95.0
88	L224*	99.8	"	100.0	"	93.2
89	L225*	79.1	"	100.0	454AS, ESL	91.4
90	L226*	81.3	"	100.0	"	94.1
91	L227*	131.4	"	100.0	"	96.9
92	L228*	133.3	"	100.0	"	93.7
93	L230*	133.4	"	100.0	"	95.9
94	L232*	139.7	"	100.0	"	93.9
95	L233*	111.4	"	100.0	"	96.4
96	L234*	128.9	"	100.0	"	90.7
97	L236*	115.6	"	100.0	"	93.1
98	L237*	115.9	"	100.0	"	48.6
99	L239*	148	"	100.0	"	94.2
100	L247*	108.1	"	100.0	"	87.1
101	L248*	105.9	"	100.0	"	75.1
102	L251*	121	"	100.0	"	69.5
103	L253*	125.1	"	100.0	"	66.1
104	L256*	118.8	"	100.0	"	86.2
105	L257*	134.6	"	100.0	"	90.2
106	L265*	147.9	"	100.0	"	88.7
107	L266*	148.2	"	100.0	"	89.8
108	L268*	135.7	"	100.0	"	95.7

		Mitogenomes			34 nuclear loci	
	Code	Coverage	Seq. strategy*	% reconstructed	Seq. strategy*	% SNPs genotyped
109	L269*	140.4	"	100.0	"	95.5
110	L277*	147	"	100.0	"	76.7
111	L279*	146.8	"	100.0	"	93.2
112	L282*	120	"	100.0	"	76.9
113	L293*	137.6	"	100.0	"	82.4
114	L294*	136.3	"	100.0	"	83.9
115	L295*	141.8	"	100.0	"	72.2
116	L296*	155.4	"	100.0	"	49.5
117	L479*	128.2	"	100.0	"	94.3
118	L480*	121.1	"	100.0	"	89.8
119	L483*	127.1	"	100.0	"	90.2
120	MVZ 192235	438.3	454LR	99.4	454AS	72.9
121	NH 15	0.6	ESL	13.1	ESL, 454AS	68.3
122	NH 16	164.4	"	100.0	ESL	0.0
123	NH 1988	56	"	100.0	"	0.7
124	NH 1989	17.5	"	99.9	"	3.3
125	NH 2016	89.2	"	100.0	"	93.1
126	NH 2050	1	"	25.1	"	0.0
127	NH 2060	31.7	"	100.0	"	0.4
128	NH 2122	88.5	"	100.0	"	3.8
129	NH 25	5.5	"	90.7	"	12.9
130	NH 2514	74.2	"	99.9	"	2.8
131	NH 2517	0.8	"	18.5	"	0.0
132	NH 2551	1.2	"	28.0	"	0.0
133	NH 2552	29.3	"	99.8	"	0.0
134	NH 2559	1.2	"	29.2	"	0.0
135	NH 2816	50.2	"	99.9	"	0.9
136	NH 2817	155.2	"	100.0	"	1.4
137	NH 2903	161.2	"	100.0	"	0.6
138	NH 2928	93.8	"	99.9	"	71.9
139	NH 3260	18.1	"	99.9	"	0.0
140	NH 36	14.4	"	97.7	"	3.3
141	NH 3984	68.2	"	99.9	"	1.0
142	NH 4723	75	"	100.0	"	19.4
143	NHMR 999000001085	38.1	"	99.6	"	0.5
144	USNM 104838		"	0.0	"	0.0
145	USNM 104839		"	0.0	"	0.0
146	USNM 114379	20.6	"	98.0	"	0.0
147	USNM 114381	1.7	"	39.2	"	8.4
148	USNM 121764	3.8	"	65.0	"	8.6
149	USNM 121766	1.2	"	32.2	"	0.0
150	USNM 151966	47.5	"	100.0	"	12.0
151	USNM 151971	22.2	"	99.1	"	17.2
152	USNM 199013	60.5	"	100.0	"	5.3
153	USNM 199014	22.8	"	99.9	"	9.1
154	USNM 590307	206.9	"	100.0	454AS, ESL	91.4
155	USNM 590310	146.4	"	100.0	"	69.8
156	USNM 590311	155.9	"	100.0	"	90.9
157	USNM 590312	182.5	"	100.0	"	93.5
158	USNM 590314	222.8	"	100.0	"	86.5
159	USNM 590316	242.6	"	100.0	"	96.8
160	USNM 590317	262.3	"	100.0	"	96.9
161	USNM 590318	213.3	"	100.0	"	94.2
162	USNM 590723	245.4	"	100.0	-	-
163	USNM 590724	259.8	"	100.0	ESL	88.7
164	USNM 590725	260.1	"	100.0	454AS, ESL	93.4
165	USNM 590726	-	-	-	454AS	72.4

		Mitogenomes			34 nuclear loci	
	Code	Coverage	Seq. strategy*	% reconstructed	Seq. strategy*	% SNPs genotyped
166	USNM 590727	120.8	"	100.0	454AS, ESL	97.9
167	USNM 590728	276.8	"	100.0	"	83.9
168	USNM 590729	217.9	"	100.0	ESL	95.3
169	USNM 597808	-	-	-	454AS	72.7
170	USNM 597809	248.2	"	100.0	ESL	88.5
171	ZRC 6028	0.53	"	12.7	"	0.0
172	ZRC 6403	6.8	"	94.1	"	19.4

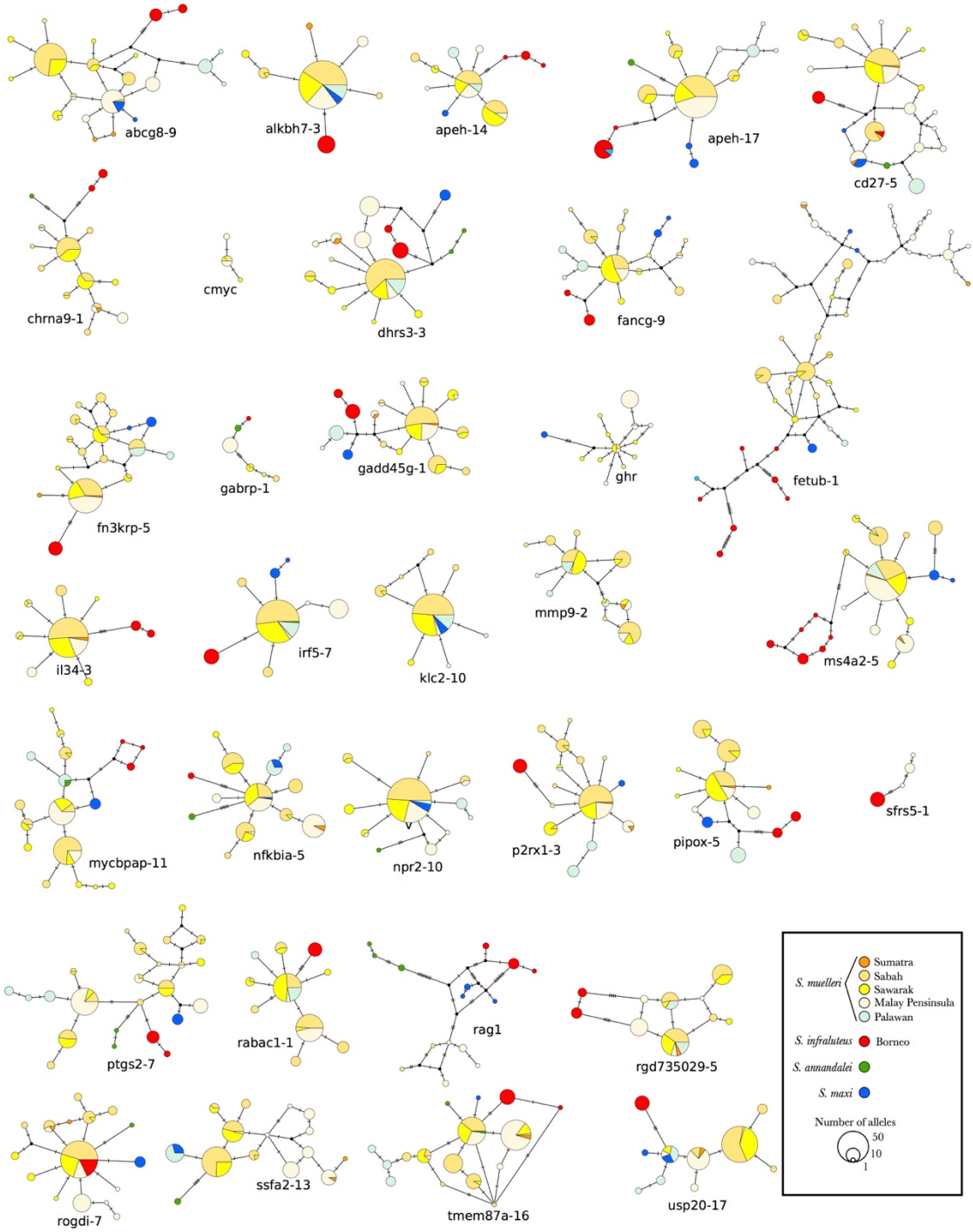
*ESL, Illumina shotgun enriched; 454LR, Roche 454 LR PCR; NESL, Illumina shotgun non enriched; 454AS, Roche 454 amplicon library.

Appendix 5.2. RAxML maximum likelihood tree with protein-coding genes of mitogenomes (all samples shown).



Appendix 5.3. TCS haplotype networks for the nuclear loci.

The size of the pie charts corresponds to the frequency of the haplotype; transversal lines represent single base pair differences or indels; small black circles represent missing haplotypes.



Conclusions

- *Lots left to learn*. Through the course of this work we have exposed cryptic and divergent diversity within Sunda rats with solid genetics, combining evolutionary inference with complete mitochondrial genomes and multilocus analysis. This, has allowed us to resolve taxonomic uncertainties in one of the most complex groups, the Rattini, including the reclassification of a species to a different genus, identify divergent lineages as new potential species, and reveal the correlation of strong genetic structure with geological past. Thus, highlighting how little is known for most mammals in Sunda, and laying methodological bases for looking into the taxonomy and diversity of similar groups.
- *High permeability of the Malacca Strait*. Species distribution (e.g. *Sundamys annandalei*), and within species patterns of diversity (e.g. *Sundamys muelleri* and *Sundamys annandalei*) support higher connectivity in the Quaternary between Sumatra and the Malay Peninsula than between these areas and Borneo and/ or Java. This supports the presence of rainforest between Sumatra and the Malay Peninsula during glacial periods when sea level was lower.
- *Borneo is a cradle of diversity.* Borneo holds largest diversity for the lowland taxa studied, while divergent lineages compared to western Sunda (e.g. *S. infraluteus*, *S. muelleri*, and *R. tiomanicus*), which support the hypothesis of the relative isolation of Borneo, likely by a large block of dry vegetation in central Sundaland during glacial periods as an ecological barrier in times when the Shelf was exposed.
- *Refuges at risk.* Species dependent on montane habitat are currently in a refugial state. The long branches upon which many of these species reside (e.g. *S. infraluteus, R. hoogerwerfi, R. korinchi*) suggests that they have been able to survive in that habitat for long periods of time. Overall, our results show the important role of tropical mountains as interglacial refugia, which promote and act as reservoirs of tropical biodiversity across glaciations. Reason for which mountain habitats should be considered as priority conservation areas in these regions. However, models based on the ecology of the most recent montane lineage (*R. baluensis*) suggest that ongoing climate change may place these mountain lineages at risk.

- *Strong selection in montane environments.* Mountain habitat seems to have driven convergence in external morphology for Sunda endemic mountain rats *Rattus korinchi*, *R. hoogerwerfi*, and *R. baluensis*. They have independently colonised mountain forest in Borneo and Sumatra, but evolved to similar external traits. While the Sumatran lineages are relatively old, *R. baluensis* stands out as a singular case of recent speciation across an ecological gradient from its lowland sister taxa *Rattus tiomanicus*. The rapid convergence in this species to the mountain morph suggests strong directional selection and/or a "island effect".
- *Bases for future research*. This work lays a strong genetic background on the neutral variation of these rats for future research on speciation and adaptation.

Acknowledgements

Over these years many are the people and institutions that have given the extensive support needed for the elaboration of this dissertation. I am immensely grateful to Jennifer for offering me the opportunity to initiate this project under her guidance. She has assisted me with all kinds of intellectual, financial and logistic backup to produce this thesis dissertation. She always had the door open for advisorship, while leaving room for me to develop own ideas, get things right and wrong, and learn. I have much affection to Jesus Maldonado who co-supervised me from Washington DC and was part of my thesis committee. He provided insightful input from the distance and when we had the few chances to meet. I must thank Kris Helgen for suggesting me to work on *Sundamys*, and for transmitting his passion and motivation for Natural History in a such a generous way. He was part my thesis committee, invited me to mammal meetings held at the NMNH, supported my research therein and to other collections. Jesus, Kris and Robert C. Fleischer kindly hosted me for several months during fruitful and enriching visits at the SCBI and NMNH, Washington DC.

Melissa Hawkins started a bit earlier her PhD. We have been going in parallel in this same trip most of the time. We survived two expeditions together in Borneo, in 2012 and 2013. I must thank her for being my partner in the field, hosting me in her house in the US, emails and guidance on lab protocols, sampling historical specimens together, and probably other things that I forgot.

I want to thank to all the extraordinary Sabahans that helped us, supported our research, and made our stay in Sabah unique. Fred Tuh, our Malaysian research counterpart, was always supportive at all levels. Our research in the field would not have been possible without him. I thank all the people that participated in the expeditions to Tambuyukon and Kinabalu in 2012 and 2013: porters, guides and field assistants from Pinawantai, Ranau. Specially thanks go to Ipe Atun, who was our field assistant in Kinabalu National Park during 2012 and 2013, and with whose family I spent several days. I acknowledge Richard, who was my guide in Tambuyukon and hosted me in his house. I thank Paco Carro and Manolo López, whose creativity and positive sense of humor in the field were greatly appreciated. To Flavia for making part of the dissertation "trip" special. All the people that participated in fieldwork in Gunung Alab and Gunung Trusmadi in 2016 deserve my acknowledgment, particularly my friend and colleague

Arlo with whom I spent two months in the field. Razak and Christopher (Pop) were excellent companions other than creative and committed field assistants in Trusmadi. Daniel Hinckley was an extraordinary person to share the time in the field with. I want to thank all porters, guides and drivers from Tambunan, rangers and staff from Sabah Parks and The Forestry Department. Specially, I would like to thank Nelly Majuakim (Sabah Parks), Ms. Chung (SaBC), Silvester (SWD), and Mr. Salleh Intang (Forestry Department), for all kinds of professional support to our research in Sabah. I also thank Benoit Goossens and the Danau Girang Field Center (Sabah) for allowing me to spend one week at their research station at the Kinabatangan river.

I want to thank Carles Vilà and Martina Carrete for tutoring me at the University Pablo de Olavide. Carles and José Castresana kindly accepted to be part of my thesis committed and provided relevant insights.

To the Conservation and Evolutionary Genetics Group, mainly for all the good times, but also for discussion and other kinds for scientific feedback and support (Alvaro, Arlo, Giovanni, Irene, Mar, Inés, Carlos, Santi, Anna, Conchi, Vicente, Jorge, Alejandro, Juanma). Anna and Irene were of great help in the lab. I also want to thank the PhD students, postdocs and PI's at the SCBI at the NMNH, Washington DC and the Doñana Biological Station. Elena Marmesat and María Lucena were fun and smart companions to discuss genomics issues. They helped me many times.

UPO and the Doñana Biological Station supported me all this years. I thank the administration for all their help. In the EBD, I thank the technicians at the Laboratory of Molecular Ecology, where I spent a large part of my time during the dissertation, and one for all to Ana Piriz. Also, the EBD collections for various kinds of support and access to the collections, and among others to Manolo Lopez and Pepe Cabot, and the LAST.

My thanks to reviewers and editors for their constructive input, specially Jake Esselstyn. This dissertation would not have been possible without the kind and professional support of worldwide museums curators and managers who opened the doors of their natural collections. I want to thank Kris Helgen, Darrin Lunde, Nicole Edminson and other mammal curators at the National Museum of Natural History, Washington DC, US; Roberto Portela from the Natural History Museum, London; Steve van der Mije and Pepijn Kamminga, Naturalis Biodiversity Center, Leiden, The Netherlands; Ted Daeschler and Ned Gilmore, Academy of Natural Sciences of Philadelphia, US; Eileen

Westwig and Neil Duncan at the American Museum of Natural History, New York, US; Larry Heaney, The Field Museum of Natural History, Chicago, US; Christopher Conroy, Museum of Vertebrate Zoology, Berkeley, US; Kees Moeliker, Natural History Museum, Rotterdam, The Netherlands; Fred Tuh, Kinabalu Park Museum, Sabah, Malaysia; Albert Lo, The Sabah Museum, Sabah, Malaysia; Chua, Raffles Museum, Singapore. I thank all collaborators, especially, Pierre Henri Fabre. I thank the SYNTHESYS Project (http://www.synthesys.info/) which is financed by the European Community Research Infrastructure Action under the FP7 Integrating Activities Program, and supported my visits to the NCB (NL-TAF-5588), Leiden, and NHM (GB-TAF-5303), London. The Spanish Ministry of Science and Innovation grants CGL2010-21524 and CGL2014-58793-P also supported this work. I was supported during four year by the Spanish Ministry of Science and Innovation Predoctoral Fellowship BES-2011-049186.

To my parents and family for their irreplaceable support in these years of ups and downs.